理想国
imaginist

# The Big Data Revolution

# 大数据

涂子沛 | 著

除了上帝，任何人都必须用数据来说话。

正在到来的数据革命、以及它如何改变政府、商业与我们的生活

0

1

GUANGXI NORMAL UNIVERSITY PRESS
广西师范大学出版社



imaginist

想象另一种可能

理
想
国
imaginist

一个真正的信息社会，首先是一个公民社会。

——作者题记

薛华成［中国管理信息系统的主要开创者、澳门科技大学行政与管理学院名誉院长］

通过历史故事的讲述，用近似新闻评论的笔法，《大数据》这本书以美国为例介绍了由经验管理到数据管理这场正在发生的革命性变革，通俗形象地讲清楚了科学的道理，实属难得。这对学者们扩大视野，对管理干部深化科学发展观的认识，都有十分重要的价值。

施伯乐［复旦大学计算机研究所所长、上海（国际）数据库研究中心主任］

《大数据》这本书揭示了数据在未来社会中的革命性作用。大数据将是下一个社会发展阶段的“石油”和“金矿”。无论是个人、企业还是国家，谁能更好地抓住数据、理解数据、分析数据，谁就能在下一波的社会竞争中脱颖而出。关于数据的知识，将成为个人知识结构中的必备要素和基础。

赵嘉敏［东西网总裁、译言网创始人之一］

子沛兄的《大数据》这本书，帮助我们形象地了解信息自由、技术创新与社会进步的关系。我最感兴趣的是美国信息自由制度建立的过程。它让我看到，一个社会、一种文化，是如何通过自省式的努力，认清比自身存在更本源的客观规律。这种制度一旦建立，就如同打开了控制某个基因群组的开关，后面所呈现的演化，虽然眼花缭乱，反而显得平淡无奇、顺理成章了。

董洁林［苏州大学企业创新和发展研究中心主任、教授］

18世纪启动的工业革命为人类创造了巨大的财富，也极大地推动了社会变革。20世纪中启动的信息革命对人类社会也将产生巨大的冲击，而这个进程刚刚开始。涂子沛先生所著的《大数据》一书，用有趣的故事带我们回顾信息时代的里程碑，使我们透过未来的变幻不清的迷雾，展望和畅想信息革命的大趋势。

汪小帆［上海交大电子信息与电气工程学院教授］

《大数据》值得仔细阅读与思考。要让数据更好地说话，既依赖于数据科学与技术的发展，更需要政府、机构和公民的数据意识的不断提高。

张晖［卡内基梅隆大学计算机系终身教授］

数据科学正在成为一个新兴的学科和领域。涂子沛先生的《大数据》这本书，为我们揭示了大数据浪潮发生的前因后果，以及这种技术进步对政治、经济、文化的种种影响。希望读者从本书中获益，以积极的态势应对中国的大数据挑战。

Ramayya Krishnan［卡内基梅隆大学海因兹学院及信息管理学院院长、终身教授］

Big Data has the potential to fundamentally transform Society. However, unlocking this potential will require careful attention to data governance and insightful application of data analytics combined with an

environment that spurs managerial innovation. Zipei's book provides a roadmap for the Chinese professionals using the American experience as context. As the dean of the Heinz College and its School of Information Systems at Carnegie Mellon University, I am very proud of what Zipei has accomplished. His book combines government, business and technology concepts in a holistic manner, embodying the unique strengths of programs and faculty of the Heinz College. I believe this book will stimulate new ideas in China to meet the big data challenge.

中文译文：大数据具有催生社会变革的能量。但释放这种能量，需要严谨的数据治理、富有洞见的数据分析，以及一个激发管理创新的环境。涂子沛的《大数据》这本书总结了美国的经验和实践，并为中国提出了一个路线图。我为他的成就感到骄傲。这本书完整地融合了技术、政治、商业三个角度，体现了我们学院师资教学的独特优势。我相信本书能够在中国激发新的思想和讨论，帮助中国迎接大数据的挑战。

涂子沛 著

# The Big Data Revolution

# 大数据

正在到来的数据革命，以及它如何改变政府、商业与我们的生活

广西师范大学出版社
· 桂林 ·

图书在版编目（CIP）数据

大数据：正在到来的数据革命，以及它如何改变政府、商业与我们的生活［2.0升级版］涂子沛著．—桂林：广西师范大学出版社，2014.4（重印）

ISBN 978-7-5495-1837-1-01

Ⅰ．①大…　Ⅱ．①涂…　Ⅲ．①信息产业—产业经济—通俗读物　Ⅳ．①F49-49

中国版本图书馆CIP数据核字（2012）第105215号

广西师范大学出版社出版发行

桂林市中华路22号　邮政编码：541001

网址：www.bbtpress.com

出 版 人：何林夏

全国新华书店经销

发行热线：010-64284815

山东临沂新华印刷物流集团有限责任公司

开本：700mm×1000mm　1/16

印张：24　字数：242千字　图表：145幅

2013年4月第2版　2014年4月第8次印刷

定价：49.90元

献给

母亲魏木兰和妻子张炜婷

感恩无私的爱

# 目　录

# 序言一　大数据：为华文世界提出一个重要话题

许倬云

涂子沛先生的新著《大数据》，已经完成，是一部300多页的大作。最近他将这本书的打样稿送来给我看，并且要求我作序文。可惜的是，我将要远行，而他索序甚急，我只能在浏览一遍之后，发抒自己一些感想。至于细细咀嚼和消化这本有趣作品的内容，必须要在一两个月以后，也就赶不上涂先生大作的出版时间。我事先声明，这些只是我浏览他大作以后的一些感想，非常立即的直觉，也一时谈不上深刻的见解。

这本书的内容，强调今天是一个大量数据公开于大众的时代。自从资讯革命以来，资讯工具、硬件、软件，平行发展、与时俱进。数据在我们生活之中，日常接触，已是处处可见的现象。收集数据，当然是靠电脑的快速分类和记录，然而更重要的是搜索引擎的进步，与网络之间网际的交流。到今天，一个一个网，不但可以串连在一起，互相沟通，而且“云端”（云计算）的设计，可以将每一个地区个别数据，储成一个大的数据库，有助于我们更迅速广泛地搜索。

这些现象，20世纪最后四分之一以来，已经呈现加速度发展的新事物，在21世纪进展速度之快，更是铺天盖地、无所不在。涂先生在这本书里强调：不仅数据经过管理而大量地存在，而且，在现代的国家，尤其是民主社会，开放的社会与政府之间，经由数据，彼此一目了然、无所隐瞒。一个自由的社会，掌握公权力的政府，跟任何其他政府一样，

有压倒社会的庞大力量，因为他们手上掌握了人生需要的许多数据。不过，这些数据，在自由的社会，公民也可以一样取得，使政府所作所为，可以摊开在天地之间，让我们检验。过去封建专制和集权各种政体，其执政者能压迫老百姓，而老百姓没有办法回制公权力的压迫。涂先生特别标榜，美国奥巴马接任以后，尽力将数据开放于大众，固然奥巴马是一个有开放心胸的政治人物，如此将政府掌握的数据，大量地开放于群众，也是拜时代之赐，有如此的机缘，才能将数据公开。

涂先生引用胡适之先生与黄仁宇先生的话。胡先生说中国人习惯于“差不多先生”，凡事马马虎虎、不求精确。黄仁宇先生认为，中国不懂得用数字来管理国家。涂先生引用这两位先生的名言，当然是要彰显传统中国和今天美国之间的巨大差异。不过我必须有所说明：胡先生和黄先生的话语，都是“爱之深而责之切”的心态，他们身经当时中国的混乱，激愤而出此感言。

从历史上看，不论中国和西方，任何国家发展到可以有一个复杂文官系统管理以后，没有不依照数据来治国的。人口、资源、土地、财产种种的统计数字，在中国历史上，自从战国时代形成列国的国家体制以后，没有一个朝代不具有一定的数据库；只是以今天的标准来讲，粗糙和细密之间，古今有很多的差别而已。以汉代为例，汉简所显示的家户统计，每一户中的人口，男女老小，以及拥有的资产数目字，都详细统计，而且不论是居延边塞，或是荆州内郡，格式一致。汉简各种家户统计，与唐代西域州府的记录对比，其内容格式也是相当一致。这种基本的数据，在列朝的会典中，都见到其大概。当然，各个朝代的数据，有做得好的，也有做得差的。大致讲起来，外族侵犯中国建立的朝代，以武装力量强制建立政权，也往往依靠暴力的掠夺，取得他们所需要的资源。一个上轨道的朝代，其数据还是相当完整。

再看西方历史。希腊时代，我们了解的资料不够。罗马帝国时代，全帝国包含各种不同的政治单位，并没有一个大一统的文官政府；因此，全国性的资料库似乎不存在。等到中古黑暗时期，国不成国，地方不过是大小封建领主占有领土而已，他们并没有建立详细的资料库。近代以来，列国各自组成完整的主权国家，这些数据也纷纷出现了。

这是以历史上政权掌握数字而言。一家大型的企业，例如，中国清代的票号，如果他们手上没有复杂的数据库，就不能进行汇兑、放款、存款等等活动。英国的东印度公司，手上握有丰富的资源，他们也不能不具有一个相当完整的数据库，否则无以经营这么复杂的开拓业务。

今天的数据时代，我已经在前面提过，不仅公司单位都有搜集数据的能力，而且更重要的，有搜寻引擎可以将资料迅速检索，从其中归纳出条理，有助于了解情况。举一个例说，最近我们才看到，数据资料显示，美国百分之一的人口，拥有全国财富百分之四十以上，百分之九十九的人口，拥有全国财富才过半而已。对于许多长期习惯于美国是开放社会的一般人民，这一组数据显示的现象，几乎可说是理想的破灭，使大家必须检讨：美国真是如此开放吗？还是相对地在逐渐关闭？是不是财富与权力，已经逐渐集中到社会顶端一小撮的人手中？他们以财富作为魔法师的指挥棒，安排了我们的生活、决定了我们的未来。这种现象能够暴露于众，当然就因为在美国究竟资讯是公开的。

相对而言，在极权的国家，他们手上拥有足够的资讯，足以利用这资讯，掌握每一个人的日常生活；“老大哥”的影子，可以无所不在。资讯时代，对于极权的掌权者，他们拥有无可比拟的强大工具，甚至于比坦克车和催泪弹更为有效。资讯管理、资讯控制，是无影无踪，又是无

所不在。说到这里，我们不能不更多警惕。

作为一个史学工作者，看惯了世间的灾难和创伤，不能不提出警告：这个中性工具，也只有在了解到资讯工具阴暗面——双刃剑的特性，在权势独占这一工具时，可能出现的危险。有此认识，我们才能善于利用这了不起的工具，开创更好的未来，也防治不虞的灾害。

涂先生这部书，清楚地叙述了资讯时代对我们生活的影响与社会的控御力。他讨论的范围方方面面、极为广泛。我盼望有了这本书作为起头，还有很多对资讯工具有研究、也有心得的人，参加讨论，让我们更清楚地了解，这个21世纪正在坐长的新的知识工具。为此，我们要对涂先生致敬与致谢，因为他为华文世界提出一个重要的话题。

2012年4月8日　于匹兹堡

# 序言二　中国的雄心应该拓展到大数据领域

## 托马斯·H·达文波特*

无论是对中国政府，还是就中国的商业组织而言，《大数据》都是一本重要的书。大数据及其分析，将会在未来10年改变几乎每一个行业的业务功能。任何一个组织，如果早一点着手大数据的工作，都可以获得明显的竞争优势，正如早期在“小数据”时代脱颖而出的竞争者一样，如第一资本金融公司、前进保险公司、万豪酒店等等。时光荏苒，现在到了抓住大数据机遇的时候了。

大数据之所以产生，是因为今天无处不在的传感器和微处理器。我们正在迈进普适计算的时代。其实，所有的机械或电子设备都可以留下数据痕迹，这些痕迹表明了它的性能、位置或状态。这些设备和使用它的人，通过互联网互相交流，又形成了另外一个庞大的数据源。当这些数据和来自其他媒体、无线或有线电话、有线电视、卫星等等来源的数据相结合的时候，更加显得庞大无比。

这些数据可以被使用，这意味着我们可以把所有的商业或组织活动都视为大数据的问题。如今的制造业，大多数机器上都已经安装有一个

或多个微处理器，已经进入了大数据的状态。消费营销行业，无数顾客的交易触点和网上点击的流量，也成了大数据的问题。谷歌甚至认为其无人驾驶汽车也是一个大数据的问题。

世界各国的政府也开始认识到，他们坐拥海量数据，这些数据都有待分析。在亚洲国家的政府，也出现了大数据战略以及基于数据分析的方案和倡议。去年，新加坡成立了德勤数据分析研究所（DAI），这个新的机构是由新加坡政府经济发展委员会资助成立的。德勤数据分析研究所的目标，就是引领政府和企业对于数据的研究和应用。新加坡政府还资助了几所大学开展大数据和数据分析的研究活动。

任何一个组织，要抓住大数据的机遇，就必须做好几个方面的工作。从技术角度来看，首先要收集并且开发特定的工具，来管理大规模并行服务器产生的结构化和非结构化数据，这些数据，可能是自己专有的，也可能来源于“云”。其次，每一个组织都需要选定分析软件，用它来挖掘数据的意义。但可能最重要的是，任何组织都需要人才来管理和分析大数据。这些人被称为“数据科学家”，他们集黑客和定量分析员的优势和特长于一身，非常短缺。聪明的领导人，将想方设法留住这类人才。

不少公司都意识到了这种难得的机遇，现在已经采取了行动。例如，通用电气将投资15亿美元在旧金山湾区建立一个全球软件和分析中心，作为其全球研发机构的一部分。这个中心拟雇用至少400名数据科学家，现在已经有180名各就其位了。通用电气在全球拥有超过1万名工程师从事软件开发和数据分析工作，通过共同的分析平台、训练、领导力培训以及创新，他们的努力得以协调合作。通用电气对于大数据的研究活动，相当一部分集中在工业产品上，例如机车、涡轮机、喷气发动机以及大型能源发电设施。

对任何一个试图通过大数据获得成功的组织来说，通用电气的投资规模和雄心都是一个榜样。在很多领域，中国政府和中国的企业都有雄心勃勃的计划，这引起了全世界的关注，这些雄心和计划，现在应该拓展到大数据的领域。涂子沛先生的这本书，将在这个重要的领域，为中国政府和企业的努力提供引导和帮助。

* 托马斯·H·达文波特（Thomas H. Davenport），哈佛大学商学院访问教授、巴布森学院（Babson College）信息技术与管理学总统杰出奖教授，2003年，他被《咨询》杂志评为全球“最优秀的25位咨询大师”之一，2005年被《优化》杂志评为世界商业与技术分析顶级三强之一。

# Foreword
［达文波特序言英文原文］

This book is an important one for Chinese government and business organizations. Big data and analytics based on it promise to change virtually every industry and business function over the next decade. Any organization that gets started early with big data can gain a significant competitive edge. Just as early analytical competitors in the“small data”era (including Capital One bank, Progressive Insurance, and Marriott hotels) moved out ahead of their competitors and built a sizable competitive edge, the time is now for firms to seize the big data opportunity.

The pervasive future of big data is enabled by the pervasive nature of sensors and microprocessors today. We are entering into the ubiquitous computing age now. Virtually every mechanical or electronic device can leave a trail that describes its performance, location, or state. These devices, and the people who use them, communicate through the Internet—which leads to another vast data source. When all these bits are combined with those from other media—wireless and wired telephony, cable, satellite, and so forth

—the future of data appears even bigger.

The availability of all this data means that virtually every business or organizational activity can be viewed as a big data problem or initiative. Manufacturing, in which most machines already have one or more microprocessors, is already a big data situation. Consumer marketing, with myriad customer touchpoints and clickstreams, is already a big data problem. Google has even described the self-driving car as a big data problem.

Governments have begun to recognize that they sit on enormous collections of data that wait to be analyzed. We can see big data and analytics initiatives among governments in Asia. Last year, Singapore helped to launch the Deloitte Analytics Institute (DAI). This new institute is sponsored in part by the Economic Development Board of the Singapore government. The DAI's goal is to do research and thought leadership on the application of analytics to government and business. Singapore has also sponsored several university-based research initiatives on analytics and big data.

Organizations that want to pursue big data opportunities need to begin working along several fronts. From a technology standpoint, they need to acquire and develop tools to manage both structured and unstructured data in massively parallel server environments, either on premise or in the cloud. They need to select analytical software to make sense of the data. Perhaps most importantly, they need to hire or develop the human talent to manage and analyze big data. These people are typically known as“data scientists”—hybrids of hacker and quantitative analyst—and they are in extremely short supply. The wise executive will develop approaches to securing the best people.

Some companies are beginning to realize the extent of the opportunity, and to act upon it now. GE, for example, has committed to spend more than $1.5 billion to develop its Global Software and Analytics Center in the San Francisco Bay Area as a part of its Global Research organization. The company plans to hire at least 400 computer and data scientists at this location, and has already hired 180. Globally GE has over 10,000 engineers engaged in developing software and analytics products and services, and their efforts will be coordinated through common analytics platforms, training and leadership education, and innovative offerings. A significant portion of big data activities at GE will be focused on industrial products, such as locomotives, turbines, jet engines, and large energy generation facilities.

The size and ambition of GE's commitment should set the tone for other organizations that want to succeed with big data. Chinese government agencies and firms are noted worldwide for their ambitious plans in other domains, and these should be extended to big data. Zipei Tu's book will help to guide government and business organization's efforts in this important area.

Thomas H. Davenport

# 序幕　新总统的第一天

你们每个人，都可以拿了毕业证、走下这个讲台，然后去追求锦衣玉食等等这个金钱社会视为理所当然的东西。你可以选择只关心自己的喜怒哀乐，把你的生活和国家的发展割裂开来。

但我不希望你这样做。这不仅仅是因为你对那些没有你幸运的人负有责任，尽管你确实负有责任；也不仅仅是因为你对帮助你走到今天的人欠有债，尽管你确实欠下了债。

这是因为：你对你自己负有使命和责任。这是因为：我们个人的命运依赖于群体的命运。这是因为：如果你仅仅考虑你自己、满足眼前的需要，这是一种贫乏。这是因为：只有你把你自己的战车和其他一些更伟大的东西绑定到一起的时候，你才能发现你真正的能量，才能发现你为美国这个国家继续书写历史时所能扮演的角色。[01]

——奥巴马，在卫斯理大学毕业典礼上的演讲，2008年5月27日

2009年1月21日，星期三。

对美国白宫而言，这是特殊的一天，因为椭圆形办公室迎来了新的主人：第44任美利坚合众国总统巴拉克·奥巴马。[02]

8点35分，这位48岁的黑人总统来到了办公室。毫无疑问，这将是非常忙碌的一天。在一个政治姿态极其重要的时代，他第一天的举措将会引起广泛的评论和解读。他也要利用这个机会，迅速为他的新政府定

下一个基调和方向。

日程安排得很满。按照传统，新总统上午要去华盛顿国家大教堂，参加就职礼拜。中午，他要分别与埃及总统穆巴拉克、以色列总理奥尔默特等四位中东领导人通话，讨论该地区的和平问题。接下来，他要参加内阁的宣誓仪式并发表讲话。2点半，他要出现在白宫开放日的现场，接见上百名来白宫参观的美国普通公民。4点整，经济顾问委员会和驻伊拉克部队的司令官要和他分别会面，讨论8000多亿的救市方案以及伊拉克撤军的具体计划。

萎靡不振的国内经济、久拖不决的海外战争，这些阴霾正笼罩着这位新的总统。

宽大锃亮的办公桌中间，静静地躺着一封信："致：第44任美国总统"。

这是前总统布什留下的。在告别白宫之际，给下一任总统留下一封信，已经是美国上百年的传统了。

布什在信中陈述了当前美国面临的种种艰难形势和挑战，最后祝他好运。读完这封信，奥巴马抬头望向窗外，陷入了短暂的沉思。

# 一人一票：把“黑”人送进“白”宫

同胞们，对我而言，这是一个惊喜的时刻。想想看，当年那个来自奥地利的孩子，他骨瘦如柴，今天却成为加利福尼亚州的州长，并站在麦迪逊花园广场上，代表美国总统讲话。这是一个移民的梦想，是一个美国梦！[03]

——阿诺德·施瓦辛格

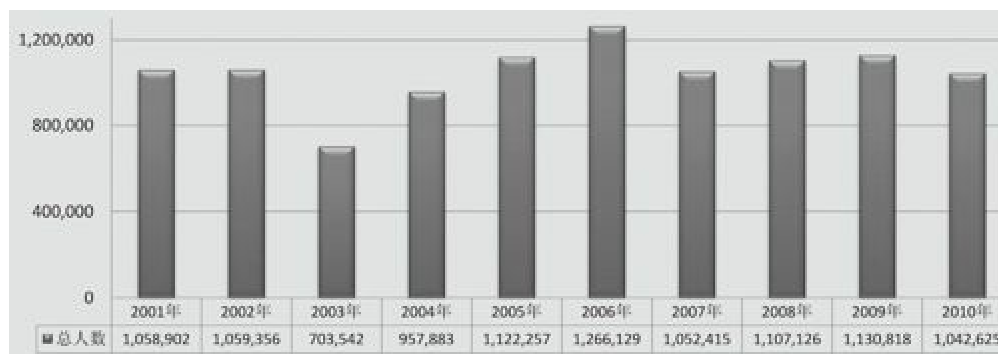
在共和党2004年全国大会上的演讲，2004年8月31日

奥巴马的故事，是一个“美国梦”的故事。

带着对财富的追求、对平等的向往，每年都有成千上万的人涌进美国这片土地。上百年来，各式各样、不同版本的“美国梦”在这里生长孕育、传承演绎。

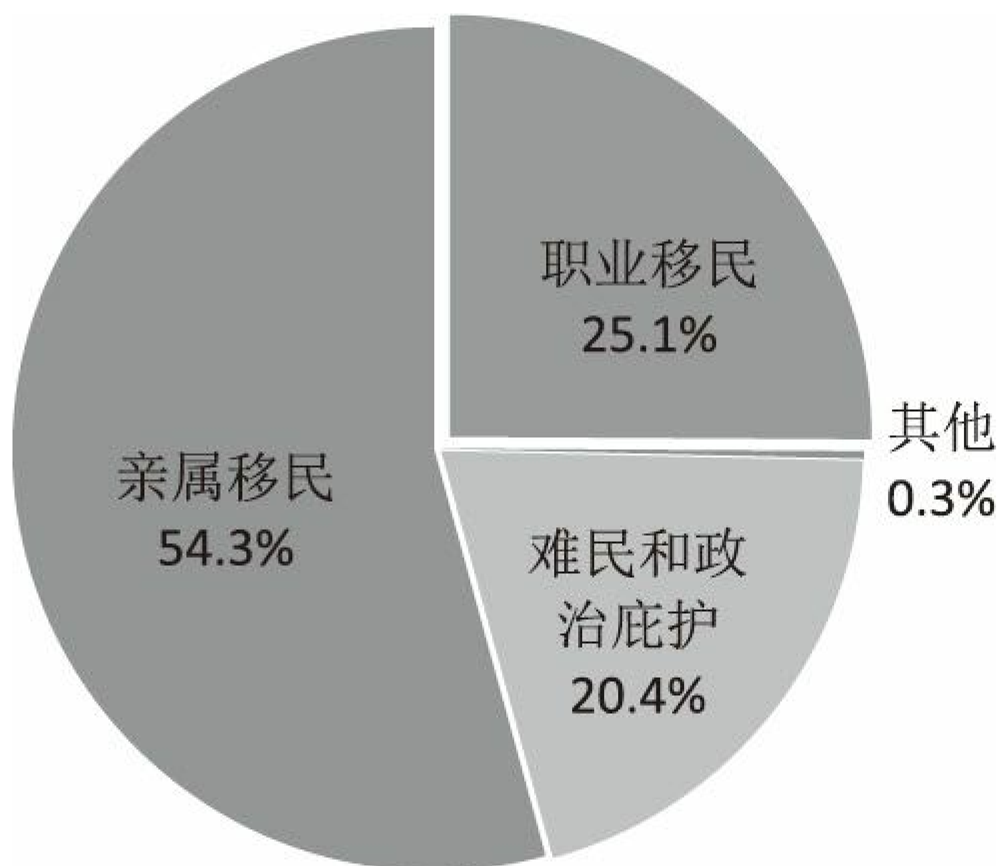
只要通过自己的不懈努力，即使是外来的移民，也可能获得巨大的成功。这是“美国梦”中最炫目的一环。

**2001**年至**2010**年在美国申请并获得居留权的移民人数



说明：近十年，平均每年有105万人获得美国的居留权（实际每年的申请人数都大大超出获准人数），这105万人来自全世界200多个国家和地区。（数据来源：美国国土安全部）

移民美国的中国人构成




说明：近十年来，每年平均有7万以上中国人（不包括中国台湾）获得美国的居留权。2005年起，中国已经超过印度，成为美国海外移民的第二大来源国，第一大来源国是墨西哥。以2010年为例，该年度共有73438名中国人成功移民美国（不包括中国台湾），其中亲属关系的移民占一半以上。（数据来源：美国国土安全部）

1968年，21岁的施瓦辛格从奥地利来到美国。1983年，他加入美国

国籍，成为美国公民。2003年，56岁的他当选为加利福尼亚州的州长。

和施瓦辛格相比，奥巴马一出生就是美国人，但他是黑人。他1961年出生的时候，美国的黑人还没有真正的投票权。[04]他的父亲，只是一个从肯尼亚来美国短暂学习的留学生，连移民也谈不上。

奥巴马和布什也不一样。和布什相比，奥巴马的成长和崛起，可以说是完全另类。这种另类，大部分美国人都认为，又是一个美国梦的生动注解。

布什可谓是衔着“金钥匙”出生的。他生于政治世家，其祖父做过美国国会的参议员，父亲担任过美国总统。布什从耶鲁大学毕业后，先从军、后经商，经州长、再总统，走的是一条标准的精英道路。

而奥巴马的父亲来自非洲东部的一个贫穷小国：肯尼亚。老奥巴马是一个牧羊人家庭的孩子。1960年，他历尽艰苦、远涉重洋，来到美国夏威夷留学，不久后和美国女性安·邓纳姆（Ann Dunham）恋爱结婚，次年生下了黑白混血的奥巴马。

没过多久，老奥巴马就抛妻离子，回到了肯尼亚。邓纳姆也再婚并迁居到亚洲的印度尼西亚，小奥巴马在印尼接受教育，一直到10岁才回到美国，此后长期寄居在外祖父家里。

年轻的奥巴马一直在迷茫中长大，他找不到自己的身份认同。他曾经在自传中提道：作为一个美国人，他不了解美国文化；作为一名黑人，他又似乎不属于那个群体。这个问题困扰了他很久。

他刚上大学时无所事事，找不到自己人生的方向，甚至借助毒品麻醉自己。后来由于他母亲和师长的影响，奥巴马开始关心公共事务。

1985年，他从哥伦比亚大学毕业，做出了一个与众不同的决定：加入一个和教会相关的草根组织，为芝加哥南部的贫困社区工作。当年美国的年平均工资约为14500美元，奥巴马作为常春藤名校的毕业生，却接受了12000美元的年薪。

奥巴马在那里工作了3年。他看到了贫穷，体会到了无奈，学会了倾听、斡旋，积累了丰富的草根经验。这段经历，成了他人生的转折点。

在他竞选总统期间，2008年5月，奥巴马受邀在卫斯理大学（Wesleyan University）的毕业典礼上发表演讲。他又回顾了这段经历。他说，没有别人为他提供的帮助，他走不到今天；但没有他自己试图去帮助别人、为社区提供服务的过程，他就更不可能找到人生的方向。他的结论是，个人的生活和公众的生活其实是不可分割的，个人的救赎取决于集体的救赎。一个国家需要改变，就需要有更多的人站出来，承担起更多的公共服务和公共责任。

他就是这么做的。

1991年，奥巴马从哈佛大学取得了法学博士学位。5年之后，他参加了伊利诺伊州参议员的竞选；作为第13选区（芝加哥）的代表，他当选为州议会参议员。其后，奥巴马在这个位置上连选连任。2000年，他参加了联邦众议员的竞选，但以失败告终。2002年，他决定竞选联邦参议员。这次选举活动，创下了几个美国之最，奥巴马很快声名鹊起，成了全美超具人气的政治明星。

选战在2002年7月打响。共和党、民主党的两党初选（Primary），共有16名候选人，其中有7名是百万富翁，竞争非常激烈。这场初选共

耗资4600万美元，成为美国历史上最昂贵的联邦参议员初选。奥巴马在民主党的8名候选人当中以很大优势胜出，成了民主党的正式提名人。

之后，奥巴马遇到了共和党的候选人杰克·瑞安（Jack Ryan）。和布什一样，瑞安也是主流社会的精英代表。他也是哈佛大学毕业的博士，曾是一名成功的银行家、投资人，成为百万富翁之后，弃商从政，披荆斩棘，赢得了共和党8名候选人当中的初选。

奥巴马、瑞安两人的民意调查最初不相上下。但瑞安离过婚，他的前妻洁丽·瑞安（Jeri Ryan），是位影视明星，曾担纲《律政狂鲨》（*Shark*）的女主角。随着选战的步步深入，奥巴马的支持者要求瑞安公布自己全部的离婚文件，瑞安曾经一度拒绝，最后官司打到了法院，法院裁决：作为参选的政治人物，他不能享有和常人一样的隐私权，文件必须公开。

文件一公开，舆论立刻一片哗然。他的前妻洁丽向法官反映，瑞安曾经在纽约、巴黎等地的色情俱乐部要求和她发生性行为，那里面堆满了皮鞭、笼子等怪诞的器具。这段话，被记录在有关子女抚养争议的文件当中。

瑞安的名声随即一落千丈，被迫自动退选。他事后抱怨说，这是美国历史上首宗因为子女抚养文件而退选的案例。

瑞安折戟之后，共和党走马换将，又推出了资深外交家艾伦·基斯（Alan Keyes）和奥巴马对垒。和奥巴马一样，基斯是黑人，这场选举又成为美国历史上首场“黑”“黑”对决，引起了全国的关注。基斯也是哈佛大学毕业的博士，曾经担任过里根政府的助理国务卿、驻联合国大使，还参加过几届总统的大选。除了这些背景和经历，基斯本人也是善

于制造新闻的人物，曾经和现任国务卿希拉里·克林顿大打口水仗。

但仓促上阵，基斯最后以大比分的落差败给了奥巴马。

随着这场高潮迭起的选举，奥巴马开始名闻全国。

也直到这个时候，他才进入了布什的视野。在一次白宫活动上，布什碰到了来自伊利诺伊州的众议员简·夏考斯基（Jan Schakowsky），夏考斯基刚刚成为奥巴马的粉丝，正佩戴着镶有“奥巴马”名字的小徽章。布什看了两眼之后，还是把“奥巴马”拼成了“奥沙马”，接着问她：“奥沙马”到底是谁？我不认识。

夏考斯基回答说：你很快就会认识他。

夏考斯基没错。此后不久，奥巴马就完成了从芝加哥到华盛顿的跨越。2005年1月，他宣誓就任联邦参议员。短短两年之后，2007年2月，奥巴马就宣布参选总统。

2008年的总统大选，更是前所未有。竞争主要在一个女人（希拉里·克林顿）、一个老人（约翰·麦凯恩，时年72岁，是美国历史上年龄最大的候选人之一）和一个黑人之间展开。美国选民最后硬是一人一票，把“黑”人送进了“白”宫。

2009年，当选总统之后，第一夫人米歇尔曾经透露，奥巴马最想做的，是一个称职的好父亲。可以想象，他在童年所经历的家庭变故以及青春期的身份困惑，曾经给他的成长投下过多大的阴影。

# 大国新政：阳光是最好的防腐剂

我的这届政府将致力于建设一个前所未有的开放政府。我们将共同努力，以确保取得公众的信任，建立起透明、公众参与和多方合作的制度。[05]

——奥巴马，2009年1月21日

不同的人生轨迹，自然形成了不同的政治见解。竞选期间，无论是对前总统布什的内政措施，还是外交政策，奥巴马都曾进行过大量的抨击。

放下了布什的信，奥巴马走出了办公室，他知道，是时候了——是时候开始真正的改变了。

几个星期以来，围绕他即将签署的第一个命令和法案，奥巴马的助手们展开了激烈的争论。因为第一个命令和法案的象征性意义，每一任总统都格外重视。多数幕僚认为，他的第一个命令应该是关闭位于古巴的关塔那摩监狱。虽然该监狱关押的是恐怖分子，但那里最近发生的虐囚事件令国际社会嘘声一片。如果新总统的第一个命令是关闭这个监狱，奥巴马可能会为美国迅速扭转形象、直接加分。

奥巴马也同意这是丑闻，有违美国的价值观，必须立即纠正。但他还认为，仅仅关闭并不能解决根本的问题。他的根本任务，是要改变华盛顿的运行方式，营造一种新的行政文化，重建公众对于政府机关的信任。这才是各种积弊背后的根本原因，是他需要面对的深远挑战。

在竞选总统的漫漫征途中，他的口号始终是“改变，我们需要改变！”——这个改变，他一再诠释，是政府运行方式的改变，是华盛顿的改变。他也一再向选民承诺，如果当选，要建立一个透明的、开放的政府。

他计划通过即将签署的第一个命令和文件，来彰显他的决心。

中午1∶15，他来到了紧挨着白宫的艾森豪威尔行政楼，主持了内阁的宣誓仪式，并发表讲话。

“很长时间以来，我们已经习惯了为华盛顿这个城市蒙上一层神秘的面纱。但从今天起，我们将不再沿承旧例，本届政府的基石将是法治和透明。”



上任第一天，奥巴马共签署了**5**个文件，其中最重要的是《透明和开放的政府》。左一为副总统拜登。（图片来源：**Mark Wilson/Getty Images**）

“我的这届政府将致力于建设一个前所未有的开放政府。”

“每一个联邦政府的机构和部门必须知道，本届政府将会毫无保留地支持信息公开，本届政府不会站在设法截留、隐藏信息的一方。”

“为了引领一个开放政府的新时代，面对信息，政府机关的第一反应必须是公开。这意味着我们必须坚定地公开信息，而不是等待公众来查询。所有的政府机关都应该利用最新的技术来推进信息公开，这种公开，应该是及时的。”

一片掌声中，奥巴马结束了他的讲话。

他随即伏案，用他特有的左手姿势签发他任内的第一份总统令：13489号总统令。该总统令直接推翻了布什总统于2001年9月签发的13233号总统令。那是“9·11”袭击事件发生的当月，布什总统以国家安全的名义，通过该命令限制了公众查阅总统文件的权利。奥巴马首先拿自己的权力开刀，宣布放松对于总统文件的管制。

其中的意味，自然深长。

奥巴马还签署了他的首份总统备忘案[06]：《透明和开放的政府》（Transparency and Open Government）。这个备忘案具体阐述了奥巴马的执政理念，他强调建立一个开放透明、公民参与、多方合作的政府。奥巴马还在备忘案中命令联邦政府的首席技术官要会同行政管理预算局（OMB）在120天以内制订出一个开放政府的具体行动方案，作为落实第一份总统备忘案的措施。

奥巴马签署的第二份总统备忘案是《信息自由法》。在这份备忘案中，奥巴马开篇就引用了美国著名的大法官路易斯·布兰代斯（Louis

Brandeis）的名言："阳光是最好的防腐剂。"[07]奥巴马指出，美国在1966年通过的《信息自由法》（FOIA）是对开放政府这个理念最有力的承诺和表达。他强调，面对公共信息：

"当我们有所怀疑的时候，公开是第一选择。我们不能因为公开可能会导致我们的官员陷于难堪的境地，或者会暴露我们的错误和失败，或者因为莫名的担心和害怕，而把信息打上'保密'的标签。"

一个星期之后，1月28日，奥巴马在白宫就经济形势发表了讲话，又重申了自己的信念。他说道：

"我深信布兰代斯大法官的这句话：'阳光是最好的防腐剂。'我知道，重塑政府的透明不仅能保证我们实现目标，也是政府重新赢得大众信任的最佳方法"。

奥巴马第一天签署的总统令和备忘案，立刻就在联邦政府的各个部门刮起了一阵"透明、开放和纪律"的旋风。

但令人意外的是，在信息技术领域，这份备忘案也引发了一股新的浪潮。

这股浪潮，不仅冲击到美国联邦政府的各个部门，甚至波及全世界很多个国家。

《透明和开放的政府》总统备忘案
白宫新闻办公室

2009年1月21日　立即发布

发：各行政部门和机构负责人

主题：透明和开放的政府

我的这届政府将致力于建设一个前所未有的开放政府。我们将共同努力，以确保赢得公众的信任，建立起透明、公众参与和多方合作的制度。开放有助于加强民主，提高政府的工作效率，帮助我们更好地履行职能。

政府应该是透明的。公开透明有助于建立问责制，能使公众了解政府的所作所为。联邦政府掌握和维护的信息是整个国家的财富。本届政府将根据法律和政策，采取适当的措施，以便于公众查询、获取的方式发布信息，各部门还要利用现代信息技术，将日常工作和决策的相关信息上网公示，以方便公众获取。各行政部门和机构还应征求公众的反馈，以确定哪些信息对公众最有价值。

政府应该是参与性的。公众的参与能提高政府的工作成效，改善政府决策的质量。知识无处不在，政府官员能够从公众参与中获得分散在社会各个角落的知识，并从中受益。各行政部门和机构要为全体美国人民提供更多的机会来参与政府的决策，政府将从集体的智慧和信息中受益。各部门要就如何增加和改善公众参与政府决策这个问

题，广泛征求公众的建议。

政府应该是多方协作的。多方协作让美国人民积极地参与政府工作。各行政部门和机构应利用新的工具、方法和系统，在各部门之间、各级政府之间全面协作。此外，还要与非营利组织、企业和个人进行协作。各行政部门和机构要广泛征求公众的反馈，以评估协作的效果，确定新的协作机会。

我现在命令：联邦政府的首席技术官会同美国行政管理预算局的局长以及总务管理局的局长，协调相关行政部门和机构，在120天以内形成一个《开放政府的命令》，该命令将对如何落实这个备忘案提出明确具体的意见，将由行政管理预算局的局长颁布。联邦政府中的独立机构也要执行《开放政府的命令》。

本备忘案不创设任何实质或程序上的、在法律或衡平法下任何对美国及其政府部门、机构、其他实体以及其官员、职员、代理人或其他人予以执行的权利或利益。

本备忘案要在《联邦政府公报》上公布。

巴拉克·奥巴马

注释

01 英语原文为：“You can take your diploma, walk off this stage, and chase only after the big house and the nice suits and the other things that our money culture says you should buy. You can choose to narrow your concerns and live life in a way that tries to keep your story separate from America's. But I hope you don't. Not because you have an obligation to those who are less fortunate, although I believe you do have that obligation. Not because you have a debt to all those who helped you get to where you are today, although I do believe you have that debt to pay. It's because you have an

obligation to yourself. Because our individual salvation depends on collective salvation. Because thinking only about yourself, fulfilling your immediate wants and needs, betrays a poverty of ambition. Because it's only when you hitch your wagon to something larger than yourself that you realize your true potential and discover the role that you'll play in writing the next great chapter in the American story."—Obama's commencement speech at Wesleyan University, May 27, 2008

02 关于美国总统"届"、"任"、"位"三种表达方法的区别：根据美国宪法，总统任期满4年为一届，现任总统如果在任上遇刺或病逝，副总统接任，则两人同为一届；但一人如果连续担任两届总统，仍为一任；"位"是指担任过总统的实际人数。奥巴马为美国历史上第56届、第44任、第43位总统。

03 英语原文为："My fellow Americans, this is an amazing moment for me. To think that a once scrawny boy from Austria could grow up to become Governor of California and stand in Madison Square Garden to speak on behalf of the President of the United States that is an immigrant's dream. It is the American dream."—Remarks by California Gov. Schwarzenegger to the National Republican Convention, August 31, 2004.

04 林肯总统在1863年解放黑奴之后，黑人在美国还是倍受歧视，选举权没有保障。直到1965年，在民权运动的推动下，美国最终通过了《投票权法》（The Voting Rights Act），才真正落实和保障了黑人的投票权。

05 英语原文为："My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration."—Transparency and Open Government, Presidential Memorandum for the Heads of Executive Departments and Agencies, retrieved on http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/

06 在美国，总统备忘案（Presidential Memoranda）是用来表明总统的决心和观点，以及记录历史事实的，其法律效力相当于总统令。

07 英语原文为："Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best of disinfectants; electric light the most efficient policeman."—Justice Louis Brandeis, 1913.

# 上　篇　帝国风云
# 得数据者得天下——美国的成功经验

我们信靠上帝。除了上帝，任何人都必须用数据来说话。[01]

——爱德华·戴明（1900-1993），美国管理学家、统计学家

资本主义社会，是一种现代化的社会，它能够将整个的社会以数目字管理。[02]

——黄仁宇（1918-2000），美籍华人历史学家

注释

01 英语原文为：“In God we trust; everyone else must bring data.”—Edwards Deming
02 《我对“资本主义”的认识》，黄仁宇，1986年。

# 第一章　历史争战《信息自由法》

信息之于民主，就如货币之于经济。[01]

——托马斯·杰斐逊（1743-1826），第3任美国总统

（向公众）隐瞒政府实情的力量，就是摧毁这个政府的力量。这种力量不能轻易授予，也不能轻率使用。[02]

——美国国会《信息自由委员会的报告》，1976年

在美国200多年的历史河流中，作为总统，奥巴马开放政府的雄心，其实并不多见。

但早在开国之初，美国的建国者就对信息的作用展开了思考和辩论。《独立宣言》的起草人、美国第3任总统托马斯·杰斐逊（Thomas Jefferson）说：

“信息之于民主，就如货币之于经济。”

言下之意，如果信息停止了流动，就像货币停止了流通的经济一样，民主的制度将会名存实亡。

美国早期的信息公开主要集中在立法机关：国会。从1841年起，美国国会的立法辩论、投票过程就向新闻界开放。

由于三权分立的体制，政府获得了相对独立的行政权。1789年，美

国国会制定了《管家法》（Housekeeping Act），规定了行政机关必须在统一的出版物上公开政务信息，但对于公开的内容，行政长官有自由裁定权。

没想到，这份立国之初通过的法案，后来“管”了150多年。

进入20世纪以后，美国开始步入大政府时代：政府规模不断膨胀，开支越来越大，公文数量急剧增加。为应对这种局面，国会于1935年、1946年先后通过了《联邦登记法》（Federal Register Act）和《行政程序法》（Administrative Procedure Act），创建了专门公布联邦政府信息的《联邦登记日报》（Federal Register），规定公众可以向政府提出信息公开的要求，但如果危及公共利益，政府有权拒绝。

一句话，美国行政部门的信息公开，100多年来，主要还是政府部门自己说了算。

直到上个世纪50年代，世界形势风起云涌，一场波澜壮阔的信息公开运动才在美国真正拉开了序幕。但这场运动的中流砥柱，却不是美国总统。

相反，这个时期的几任总统，在其中扮演的角色，都并不光彩。

# 第四股力量：知情权的起点

一个自由的新闻行业是居于政府和人民之间的伟大解释者。你，如果同意给它加上桎梏，无异于给自己戴上脚镣。[03]

——乔治·萨瑟兰（1862-1942），美国最高法院大法官，1936年

这场运动的源头来自于美国社会的“第四股力量”（Fourth Estate）。

在美国，第四股力量指的是除了立法、执法和司法三权之外的“新闻和报纸”。

谈到新闻和报纸在美国的作用，又不得不提到杰斐逊，他说过：

“如果由我来决定，我们是要一个有政府、没报纸的社会，还是一个有报纸、没政府的社会，我将毫不犹豫地选择后者。”

由于在立国时就确定了言论自由的原则，美国的记者和编辑一直都是推动社会发展进步最重要的一股力量。

1944年6月，英美盟军在诺曼底登陆，德国开始全面溃败。这标志着第二次世界大战即将结束。指日可待的胜利，令美国的新闻界兴奋不已，因为，这意味着解除战争时期的新闻管制，重新开启一个新闻自由的年代。

在这个历史的转折点，美国的记者、编辑和律师一起拉开了信息自由运动的序幕。他们中的不少人，也因此青史留名。

1945年1月，美联社的执行主编库珀（Kent Cooper）率先在美国提出了“知情权”的概念（Right to Know）。他在《纽约时报》发表文章说：“知情权是指人民有权知道政府的运作情况和信息。如果不尊重公民的知情权，在任何一个国家，甚至全世界，都将无政治自由可言。”[04]

美国新闻界不仅对国内的新闻自由和信息开放充满了憧憬，甚至雄心勃勃、跃跃欲试要将新闻自由推向全世界。他们相信，缺乏新闻自由，正是一些专制国家能够发动战争的原因。他们预见到，二战期间高涨的国家主义可能会成为战后世界范围内影响信息自由流通的最大障碍。为了克服这个障碍，1948年，美国报纸编辑协会（ASNE）成立了“世界信息自由委员会”（以下简称世委会）的分支机构，[05]著名的编辑沃尔特斯（Basil L. Walters）担任了该委员会的首任主席。

但谁也没有想到，第四股力量的良好愿望，最后竟然两头落空。

世委会成立之后，在联合国进行了大量的动员和游说工作。在他们的努力下，1948年，有55个国家达成了世界新闻信息自由流动的初步协议，但由于前苏联的反对和搅局，这个方案最后在联合国大会上流产。

不仅在国际战线受挫，而且在美国国内也阴霾重重。二战结束后，美国确实撤销了战时新闻审查办公室。但没过多久，东西方的冷战却开始兴起，杜鲁门总统宣布全力对抗“赤色革命”，为了“国家安全”，他曾经多次直接下达命令，封存公务记录、抵制信息公开。

作为世委会的掌舵人，沃尔特斯很快认识到：国内的新闻管制并没有真正解除，公众的知情权依然没有保障，和战争时期相比，甚至有过之而无不及。

门前雪都扫不干净，谈何世界？

1950年，东西方的冷战格局已经清晰地形成，两个阵营的对峙，如山雨欲来、一触即发。沃尔特斯审时度势，最终决定把视线收回到美国，他宣布将“世界信息自由委员会”中的“世界”两字去掉，以务实的态度，把工作重点锁定在美国本土。

沃尔特斯把这个转变称为“再次觉醒”。他重新定义了信息自由委员会的目标和使命：

“我们的工作就是要让人民充分意识到以下事实：政府所有的公共记录都属于人民，公务员仅仅是人民的服务人员，而报纸，好比人民的眼睛，帮助人民把公共记录和政府官员置于阳光之下。”[06]

沃尔特斯后来还当选为美国报纸编辑协会的主席。他在信息自由委员会的两位继任者，帕勃（James S. Pope）和韦金斯（J. Russell Wiggins），都为信息自由运动做出了杰出的贡献，也先后担任了美国报纸编辑协会的主席。韦金斯曾经对美国报纸编辑协会的使命有过一个更直接的描述和定位，他写道：“我们的协会必须挺身捍卫知情权，和限制信息自由的做法斗争。”

1951年，杜鲁门总统签发第10290号行政命令，在美国历史上第一次把军事机构的保密制度引入了普通的行政部门，他授权联邦政府的部门首脑可以对公共文件进行分级保密管理。只要定为保密的文件，公众一律无权查看，一时间，“保密”之风在政府各个部门蔓延。

这个时候，帕勃正担任信息自由委员会的主席。针对杜鲁门的做法，他认为必须从法理上厘清问题，找出信息自由的依据，遂委托报纸编辑协会的律师哈罗德·克劳斯（Harold Cross）对美国各级政府的信息使用管理情况进行调研。

克劳斯是美国著名的律师，也是哥伦比亚大学的新闻学教授。1953年，他在两年调研的基础上，出版了《人民的知情权》（*The People's Right to Know*）一书。

开篇明义，克劳斯就提出：

“公共事务，就是大家的事。人民有权利知道这些事。如果没有知情权，即使在一个民主制度下，人民所能做的也不过仅仅就是投票改选他们的国王罢了。”[07]

克劳斯认为：只有拥有信息自由，人民才能真正拥有对公共事务的发言权。但美国的“知情权”，并没有明确的法律保障：所有行政部门的档案和记录，实际上都处于“准机密”的状态。因为是否属于“机密”，完全取决于当权领导一时的态度。

克罗斯总结说：既然是合法的权力，就不能寄望于当权者是否开明或宽容，也不能系于领导人的道德和品质，而必须受到法律的切实保障。他还为如何落实知情权指明了方向：要清除一大一小两个法律障碍，大的是1946年的《行政程序法》，小的是1789年的《管家法》，这两部法律都用模糊的语言规定了政府机关对其文件是否公开具有决定权。

1953年出版的这一本《人民的知情权》，为美国整个信息自由运动指明了方向、奠定了框架，被后世誉为信息自由运动的“圣经”。后来美国出台的《信息自由法》，基本主张也都来源于此。

# 国会议员：孤独的战争

那美好的仗我已经打过，当跑的路我已经跑尽，所信的道我已经守住。从此以后，有公义的冠冕为我存留。[08]

——保罗，《圣经·提摩太后书》

也正是1953年，另一个更重要的人，约翰·摩斯（John E. Moss），走进了美国首都华盛顿。



约翰·摩斯（1915-1997）

《信息自由法》之父，担任美国国会议员25年，六朝元老、连选连任。除《信息自由法》之外，他还是美国《消费者安全法》等多部法律的制订者和推动者。（图片来源：First Amendment Center）

摩斯曾经是一个穷苦小子，名不见经传，但却被后世称为《信息自由法》之父。

1915年，摩斯出生于一个矿工家庭。母亲早逝，父亲弃家出走，12岁的他和弟弟相依为命。为了谋生，他卖过轮胎，做过股票交易员，开过运送尸体的灵车，因为筹不齐学费，社区大学都没有读完。但这样一个人，最后居然选择了从政，他加入了民主党，在1949年当选为加州立法会议员。

1953年，摩斯在加州立法会的表现获得了选民的广泛认可，又当选为美国国会的众议员。

这一年，在二战中立下了赫赫军功的艾森豪威尔（Dwight Eisenhower）完成了他从将军到总统的转变，也开始了他的白宫生涯。

艾森豪威尔是共和党员，但他延续了民主党总统杜鲁门的冷战思维。为了抵制“赤色”思潮，这两位总统先后在联邦政府解雇了2800多名有“共产主义倾向”的工作人员。作为国会议员，摩斯为此大鸣不平，他要求调阅这2800人的名册和相关档案。

这时候的联邦政府，已经是水泼不进，立即以“国家安全”的理由拒绝了摩斯的要求。摩斯认为这既不公平，也不合理，他开始在国会倡导信息公开。1955年，在他的推动下，众议院的政府运行委员会成立了政府信息分委会，摩斯担任了该分委会的主席。不久后，他就提出了《信息自由法》的草案。

一朝一野，摩斯和以美国报纸编辑协会为代表的新闻界两相呼应，形成了一股力量。

但这股力量还是屡屡碰壁。

按照克劳斯确定的方向，摩斯开始着手清除大小两个法律障碍。从

1955年到1959年，政府信息分委会共举行了173场听证会，撰写了17卷书面记录和14卷调查报告。在该分委会的推动下，1958年，国会对1789年的《管家法》进行了重新解释，明确规定禁止利用《管家法》来剥夺公众查阅政府文献的权利，这清除了克劳斯指出的“小障碍”。但因为共和党当政，国会里面没有一名共和党议员愿意联署摩斯的提案，《信息自由法》的草案长期在委员会搁浅，根本进入不了投票辩论的议程。

艾森豪威尔连任两届总统。等到1961年，肯尼迪上台，民主党也扭转了乾坤，成了国会的多数党。肯尼迪以开明著称，摩斯对此满抱希望。但这位美国历史上最受欢迎的总统，仍然不喜欢信息公开，他明确表示，大家应当自律，服从国家安全的大局。

但摩斯却固执己见，仍然在国会高调鼓吹“信息自由”。一般而言，国会议员不会和本党的总统以及本党的议会主席唱反调，但摩斯却针锋相对，紧紧抓住《信息自由法》的提案不放。可以想象他当时在国会受到的孤立，后人甚至评论说，这也是摩斯从政几十年最后终老于国会议员位置之上的根本原因。

美国国会的立法步骤




说明：这是众议院的立法步骤，参议院的步骤类似。

美国的国会由参议院、众议院组成。每个议院都按功能划分为多个委员会，委员会再下设分委会。美国国会共有200多个委员会和分委

会。每一个议员都属于一个或多个委员会或分委会。无论哪个议员提出的提案，都要在两院投票，半数以上通过之后，才能提交总统签署；总统签署之后，提案才正式成为法律。总统如果拒绝签署，提案将返回国会，重新修改、投票，或通过，或搁置流产。

1963年，肯尼迪遇刺，副总统约翰逊继位。约翰逊也是民主党员，这位新人一时擦亮过摩斯的希望火花。但他很快又失望了。和肯尼迪相比，约翰逊更烦记者对政府部门指手画脚、说三道四。他甚至直接传话给国会，即使他们通过了这个法案，他也会行使总统的搁置否决权（Pocket Veto）。多年以后，他的新闻秘书莫耶斯（Bill Moyers）回忆说，约翰逊听到这个法案的第一反应是：摩斯想干什么？他想搞砸我这届政府吗？

摩斯的提案似乎走入了历史的死角。

但历史的转变，常常因为一些偶然的事件而触发、完成。

真正的破冰之人，居然是一名来自共和党的新议员：拉姆斯菲尔德（Donald Rumsfeld）。

和出身贫穷、大学都没有毕业的摩斯不同，年青的拉姆斯菲尔德出身富裕，毕业于普林斯顿大学，他信奉自由经济学，反对政府管制，因此热情支持《信息自由法》的提案。他在1962年当选国会议员之后，批评约翰逊政府“不断滑向新闻管制、打压人民应该拥有的信息自由”，成为高调支持《信息自由法》的第一位共和党议员。他率先倒戈之后，越来越多的共和党议员联署了摩斯的《信息自由法》提案。当然，这背后更深刻的原因在于，这时候的共和党失去了政权，需要重新打开局面。

多年后，拉姆斯菲尔德成了一位赫赫有名的人物，他担任了两届政府的国防部长，并和《信息自由法》再度结缘。

1964年，摩斯终于在国会取得了足够的支持，十年磨剑，他的提案终于走出了委员会，进入了辩论阶段。

但以总统为首的联邦政府还是强烈反对这个法案。1965年，联邦政府只有27个部门，但无一例外，全部在听证会上大声说“不”！预算分析局（Bureau of the Budget Analysis）甚至评论说：“如果这个法案得以通过，吃饱了没事干的人，为了满足好奇心，都来查询政府的信息，这将产生不堪设想的严重后果。”白宫新闻秘书莫耶斯在法案的封面留下了这样的字迹：“对！我们必须继续反对这个法案。”

# 白宫当家人：一个妥协者和机动者

所有的政府，都是建立在妥协和交换的基础之上。事实上，人类所有的收益和快乐、所有的美德、所有明智的行为无不如此。[09]

——埃德蒙·伯克（1729-1797），英国政治学家

但事情的发展，急转而下。莫耶斯反倒成了第一个转变态度的白宫官员。

1965年10月，距离摩斯首次着手《信息自由法》的调查工作已经整整12年，参议院终于高票通过了《信息自由法》。次年6月，众议院也以全票通过。这个胜利，将新闻界的士气和大众舆论推向了一个高潮。

6月22日，提案送到了白宫。

按照美国宪法，约翰逊总统必须在10个工作日内表明态度：或签署，或否决。否决的提案将退回国会，再次投票，如果参众两院再次投票的支持率都高于三分之二，那总统的否决将失效，提案无需他的签署，将自动成为法律。

“春江水暖鸭先知”。作为白宫的新闻秘书，莫耶斯每天都要面对新闻界，面对强大的舆论压力，他的态度转变得最快。他知道约翰逊即便行使否决权，也可能无力回天，于是他建议约翰逊签字接受这个既成事实。

等到第8天，仍然毫无动静。7月2日，美国报纸编辑协会给白宫发出电报催促说：“截止日期将近。特此提醒：不能因为懈怠，让这项在

国会大受欢迎的提案胎死腹中。”

莫耶斯回信说：“懈怠不是我们的习惯，谢谢提醒。”

这时候的约翰逊，正在海水和火焰中一筹莫展。他拖到了法定10天期限的最后一天，7月4日，感觉确实无力回天，才在家中黯然签署了这份文件。

约翰逊总统热衷交际，喜欢记日记，事无巨细，每天都按时记下。他一贯的做派是在签署法案时举行隆重的仪式，邀请记者拍照、电视采访，将签字笔送人留念，并在日记中写得浓墨重彩。但对《信息自由法》，他没有举行任何仪式，甚至没有在日记中留下一个字的评论。

阴差阳错，7月4日，正是美国独立日，这使得后世在庆祝美国的生日的同时，也可以纪念信息自由的胜利。

莫耶斯后来辞去了白宫新闻秘书的职务，成了一名新闻工作者。他“下海”以后，也成了《信息自由法》的公开拥护者。2002年，他接受国家公共电台（NPR）的采访，又回忆了这段历史，他谈到了约翰逊对这个法案的厌恶和无奈。莫耶斯承认，当时的联邦政府确实以国家安全的名义，掩盖了太多的秘密。他作出结论说，秘密越多，腐败越多。[10]

当然，约翰逊总统其实也心知肚明，知道《信息自由法》是历史的进步。在生米煮成熟饭之后，他很快“调整”了自己的步伐，“跟”上了历史的潮流，并开始给自己脸上“贴金”。他发表声明说：

“这部法律根植于美国一条至关重要的原则，那就是只有当人民能够在国家安全允许的范围内获得一切信息的时候，民主制度才能达到最佳的运行状态，我为美国是一个开放的社会感到深深的骄傲。”[11]

伴随着如此华丽的辞藻，1967年，《信息自由法》开始生效。美国成为继瑞典、芬兰之后，世界上第三个实施《信息自由法》的国家。

但故事却远远没有结束，历史还有逆流。

《信息自由法》实施之后，联邦政府采取了消极应对的策略：一是拖，对于查阅公共记录的要求，常常拖延不予答复；二是贵，收取明显偏高的信息查询和复印费用，复印文件每页1美元，查询费用每小时7美元；三是挡，利用“国家安全”的豁免条款作为挡箭牌，由于该法规定的保密事项过于宽泛，中央情报局几乎拒绝了所有的信息查询及公开的申请。

因为这种种障碍，《信息自由法》几乎就成了一张“空头支票”。

1972年起，摩斯又重新启动了该法案的调查工作。在他的推动下，美国国会提出了《信息自由法修正案》，对信息公开的范围、时限和查询的费用作出了更细致、更透明的规定；此外，该提案还建议，如果政府以保密的要求拒绝查询，公民可以提起司法诉讼，法院拥有是否公开的最终裁判权。

这个修改，直接击中了联邦政府的“七寸”。这时候白宫的当家人，是福特总统（Gerald Ford）。

理解美国信息公开的两个重要概念：公开和发布

要理解美国的信息公开，首先要理解两个基本的概念：一是公开（Disclosure），公开是指政府对于公民或社会组织查询信息的特定要求，给予一对一的回应，将信息公开给查询方，查询方收到信息之

后，有自由使用、进一步公开的权利；二是发布（Distribution），发布是指将信息或者文件登载在政府出版物或政府网站上，广而告之。

例如，2009年11月，奥巴马第一次访问中国，他一共带了多少随从、开支了多少钱，这些信息，并不属于发布的范围，但如果有人查询，就属于公开的范围。

也可以说，公开，是政府和某一社会特定主体的关系，是点对点的；而发布，是政府和社会的关系，是点对面的。

1966年，当《信息自由法》在国会投票表决时，福特担任国会议员，他投下了赞成票。但时过境迁、位转人移，这时候的福特，作为总统，听到这个消息，心急火燎，立刻给国会发函，明确表示反对该修正案。更具讽刺意味的是，在60年代曾为《信息自由法》的通过有“破冰”之功的拉姆斯菲尔德，这时候正担任炙手可热的白宫办公厅主任一职，他也力主福特否决《信息自由法修正案》。

和约翰逊总统如出一辙，在参众两院已经高票通过的情况下，福特总统犹豫难决，也在第10天才做出决定。

不同的是，福特走得更远，最后孤注一掷，行使了总统否决权。

但有惊无险，参众两院又以三分之二的多数推翻了福特的否决。无需福特的签署，《信息自由法修正案》自动成为法律。

历史这才真正翻开了新的一页。

1975年，《信息自由法修正案》生效。此后，援引该法向联邦政府部门申请信息公开的人数大幅增加。当年就收到2万多起申请。这个数

字不断上升，2010年前后，仅联邦政府每年就会收到50多万宗信息公开的申请，《信息自由法》已经成为美国社会不可或缺的一项基本法案。

从此，美国的信息自由获得了前所未有的加速度。

政府的文件，是静态的决策记录；而会议，才是动态的决策过程。1976年，美国国会又通过了《阳光政府法》（Government in the Sunshine Act），规定除了10种涉及国家安全和公民隐私的会议之外，合议制机关的会议应该公开。[12]此后，美国公民获得了旁听政府部门会议的权利。

1978年，摩斯为他25年的国会生涯画上了句号。当这位六朝元老退休的时候，《华盛顿邮报》评论说，他离开了华盛顿，但他却给美国人的生活、政府的运行方式留下了一个深深的烙印。1997年，摩斯的生命走到了终点。这位风烛残年的老人对历史投下了最后的一瞥，他回顾说：

“我们所做的，只是一个开始。在面对挑战的时候，我没有退缩。很多时候我们不能退缩，必须战斗。那最美好的仗，我已经打过。”

和摩斯相比，拉姆斯菲尔德无疑长袖善舞，他也因此步步高升，受到几任总统的重用。1975年，作为福特身边的红人，他由白宫办公厅主任晋升为国防部长。2001年，他再次担任了布什政府的国防部长，可谓政坛“不倒翁”。“9·11”之后，以反恐战争为名，拉姆斯菲尔德又支持布什总统大幅收紧了信息公开的力度，再次成为反对信息自由的中坚人物。2006年，他也退出了历史舞台。当然，他并没有忘记自己的年轻岁月和历史成就。2010年，这位80岁的老人，还在自己的微博上喃喃自语：1966年，我也是《信息自由法》的共同发起人。[13]

拉姆斯菲尔德、福特、约翰逊和莫耶斯，他们都是受过良好教育的社会精英，面对同一部《信息自由法》，尽管心知肚明这是历史的进步，但行为态度为什么出尔反尔、前后矛盾呢？

沿着近现代政治学家、经济学家、心理学家所做的研究和探索，我们其实不难找到一个逻辑链条非常清楚的答案。

18世纪，英国著名的政治学家埃德蒙·伯克曾经一针见血地指出：

“所有的政府，都是建立在妥协和交换的基础之上。”

他认为，政治的行为，在本质上，是一种妥协和交换。人类政府的历史，其实是一部各种利益不断妥协与交换的历史，也只有在不断的妥协与交换中，各种利益才可能获得相对的平衡。他继而指出，其实“人类所有的收益和快乐、所有的美德、所有明智的行为”，无一不是妥协和交换。

对于这种妥协，美国的总统们并不忌讳。艾森豪威尔就曾经坦言说：

“一提到中间路线，人们就认为这是不可以接受的。事实上，人类所有的问题，除了道德，都处在中间的灰色地带。事情并不是非黑即白。我们总是要妥协的。马路的中间才是全部可以使用的区域，最左边和最右边，都是沟渠等危险地带。”[14]

马斯洛的需求层次金字塔




个人对“道德和真理”的追求只是其“自我实现”需求中的一小部分，很难在人的决策行为当中起到主导作用。

约翰逊在谈到他作为总统的体会时，也曾亲口承认：

“我是一个妥协者和机动者，我这样争取做成一些事情，我们的体系正是这样运行的。”[15]

那么，为什么必须妥协？其背后交换的原则又是什么？

几乎和伯克同时，经济学的鼻祖亚当·斯密在其经典著作《国富论》中提出了“理性经济人”的假设。他主张，人都是利己的，在面临两种以上选择时，总会选择对自己更有利的方案，争取最大的经济利益。

1943年，美国心理学家马斯洛提出了人类的需求层次理论，对人类如何做出妥协和交换作出了更细致的解释。马斯洛把人类的需求划分为

生理、安全、社交、尊重和自我实现五大类，这五大类需求如阶梯一样从低到高展开，低层次的需求相对满足了，更高层次的需求就会成为个人行为的主导。在人生的不同阶段和境况，会有不同的“主导需求”。最高层次的需求，是“自我实现”。所谓的“自我实现”，又分为“名利、成就、道德、真理”等等，其中，个人对名利和成就的追求又占了主导。

一切的妥协和交换，都是围绕个人的“主导需求”发生的。但恰恰在所有能成为“主导需求”的因素当中，道德是最为薄弱的，很难成为一个人行为的“主导”。

福特、约翰逊、莫耶斯和拉姆斯菲尔德的变化，就是因为人的位置变了，即使内心的道德判断还是一样，但利益结构变了，主导需求也变了，所以行为和选择也就变了。

人，绝不是天使。任何人的历史，都是一部道德、理性与私欲的斗争史；这种斗争，此起彼伏，至死方休。精英也好，领袖也罢，其实都是凡人，很多时候，都无法战胜自己的私心和欲望，需要外力的制约和推动。

约翰逊、福特就是典型的例子。即使明知信息自由是历史的进步，也拒绝签署。

但我们看到，虽然贵为总统，他们在信息自由的大潮面前，也没能成为阻挡历史进步的决定性因素。在外力的制约下，他们被迫妥协。这个外力，就是分权和制衡的制度。

# 政府VS.社会：旧剧情重现新时代

历史的教训很少被演员们自己吸取。[16]

——詹姆斯·加菲尔德，第20任美国总统

摩斯说的没错：他们那代人所做的，只是一个开始。

1946年，联邦政府拥有了第一台电子计算机。其后，计算机的数量不断增加。1955年，当摩斯第一次提出《信息自由法》草案时，联邦政府拥有45台计算机；1972年，当国会着手修订《信息自由法》时，联邦政府的计算机已经增加到6000多台。

但对于存储在计算机内部的数据信息是否属于公开的范围，1966年的《信息自由法》及1974年通过的修正案，都只字未提。

随着计算机的不断普及，法律和现实之间的脱节越来越明显，真空越来越大。联邦政府和社会各界之间开始出现了司法争议。

1976年，一家名叫SDC Development Corporations的公司（以下简称SDC）要求美国卫生、教育和福利部（Department of Health, Education and Welfare）公开《医疗文献索引系统》中保存的全部数据，遭到了该部的拒绝。SDC根据《信息自由法》将美国卫生、教育和福利部告上了联邦地区法院；一审败诉后，SDC又上诉到巡回法院，第九巡回法院最后认定该电子数据不属于公开的范围，SDC最终以败诉收场。[17]

除了公司之外，挑战者也不乏普通的个人。1984年，一位叫Dismukes的普通公民向内务部查询全国油田、天然气的出租名单，内务

部以微缩胶卷的形式提供了名单，但Dismukes坚持索要数据库格式的电子文档，内务部认为没有这个义务，拒绝提供。Dismukes也将内务部告上了法院，联邦地区法院最后认定：《信息自由法》没有规定信息提供的格式，内务部履行了义务，Dismukes败诉。[18]

当然，也有个别的幸运者，在赔上了足够的财力和时间之后，如愿以偿。考克斯报业（Cox Newspapers）要求联邦政府总务局（GSA）公开政府拥有的非军用飞行器的总数、机型等明细，该局提供了上千页的打印信息，考克斯报业坚持索要电子版的数据，最后也将总务局告上了法庭。一番纠缠之后，双方在庭内达成和解，总务局最终向考克斯提供了电子版的数据。

类似的官司频频出现。

1985年，国会对于政府机关是否应该公开电子数据召开了第一次听证会。听证会的结论是：由于无法可依，行政机关对于其电子数据的控制，近乎垄断。

但即便事实清楚、道理浅显，国会也无法立刻除旧布新。又是一番兜兜转转，终于在1991年11月，民主党参议员帕利希（Patrick Leahy）提出了《电子信息自由法》，要求联邦政府将存储在计算机内的数据信息与资料柜中的文件一视同仁，纳入公开的范围。

长江后浪推前浪，世上新人换旧人。1990年代，约翰逊已经作古，摩斯已经退休，拉姆斯菲尔德也离开了白宫，但随后发生的事情却惊人的相似：就像50年代全部的联邦政府机构反对《信息自由法》一样，绝大部分的联邦机构对《电子信息自由法》表示了强烈的反对，一致声称如果将信息自由扩大到电子数据，信息查询的要求和费用都将大幅攀

升，行政部门将无法承受。还有机构提出，他们没有义务去改造自己的信息系统，来满足公众的查询要求。

历史的进步总是一波三折。

又经过5年的听证和谈判，1996年，美国国会通过了《电子信息自由法》[19]。该法规定，所有政府数据库的电子记录都属于信息公开的范围，行政部门必须按照信息查询方要求的格式提供信息。此外，该法还建议联邦政府将公民经常查询的信息发布在互联网上，以减少公民重复查询产生的负担。

2007年，美国国会又通过了《开放政府法》（Open Government Act of 2007），明确了联邦政府信息公开的范围：不仅仅是其本身收集的信息，还包括政府委托私营机构、非营利组织收集的信息。该法还规定，如果行政部门不能在法定的时间内提供信息，则不能收取费用；司法部作为联邦政府各部门落实《信息自由法》的监督机关，其部长必须每年向国会报告有多少信息公开的要求被拒绝，并说明拒绝的原因。

从1955年摩斯着手信息自由的立法工作以来，经过了半个多世纪，一系列的法律层层叠加、不断完善，美国的信息自由最终形成了一个体系。这个体系，规定得很细致，考虑得也比较周全。例如，它甚至注意到了信息的可分割性，它规定：一份文件需要保密，并不是其中所有的信息都需要保密，凡是可以从保密信息中抽离出来的非保密信息，都应提供给查阅信息的普通公民。此外，即使是居住在美国的外国人，也可以像美国公民一样向政府查阅、索要相关的文件和信息。

这个体系，也成了世界上许多国家制定《信息自由法》的模板。截至2006年，全世界已经有60多个国家制定、实施了类似的《信息自由

法》。

本书还将在后续章节，陆续介绍更多关于公民或组织向联邦政府提请信息公开的故事和案例。

## 注释

01 英语原文为："Information is the currency of democracy."—Thomas Jefferson

02 英语原文为："The power to withhold the facts of government is the power to destroy that government. Such power is not to be lightly granted nor recklessly used."— U.S. House of Representatives Committee Report on Freedom of Information, 1976

03 英语原文为："A free press stands as one of the great interpreters between the government and the people. To allow it to be fettered is to fetter ourselves."—Justice George Sutherland, 1936

04 英语原文为："There cannot be political freedom in one country, or in the world, without respect for the right to know."—*New York Times*, Jan. 23, 1945, at 18.

05 美国报纸编辑协会为American Society of Newspaper Editors，世界信息自由委员会为World Freedom Information Committee。

06 英语原文为："Our job is to make the people of this nation conscious of the fact that all public records belong to the people; that officials are merely the servants of the people; that newspapers are the eyes of the people, keeping the eternal spotlight on officials and on public records. Unless we do this then the freedoms will slip out of our hands by default."—American Society of Newspaper Editors, Newspapers Awake to Constant Threat to Press Freedom, The Bulletin, Feb. 1, 1950

07 英语原文为："Public business is the public's business. The people have the right to know. Freedom of information is their just heritage. Without that the citizens of a democracy have but changed their kings."

08 英语原文为："I have fought the good fight, I have finished the race, I have kept the faith. Now there is in store for me the crown of righteousness."—2 Timothy 4:7, *the Bible*

09 英语原文为："All government, indeed every human benefit and enjoyment, every virtue, and every prudent act, is founded on compromise and barter."—Edmund Burke

10 Bill Moyers on the Freedom of Information Act, Public Broadcasting Service，April 5, 2002. http://www.pbs.org/now/commentary/moyers4.html.

11 英语原文为："This legislation springs from one of our most essential principles: A democracy works best when the people have all the information that the security of the Nation permits. … I signed this measure with a deep sense of pride that United States is an open society."

12 合议制是指集体领导的体制，联邦政府约有50个机关实行合议制。

13 Donald Rumsfeld (@RumsfeldOffice),"I was a co-sponsor of FOIA in 1966."Twitter, Nov 30, 2010.

14 英语原文为："People talk about the middle of the road as though it were unacceptable. Actually, all human problems, excepting morals, come into the gray areas. Things are not all black and white. There have to be compromises. The middle of the road is all of the usable surface. The extremes, right and left, are in the gutters."—Dwight Eisenhower

15 英语原文为："I am a compromiser and maneuverer. I try to get something. That's the way our system works."—Lyndon B. Johnson

16 英语原文为：“The lesson of history is rarely learned by the actors themselves.”—James A. Garfield

17 SDC Development Corporation v. F. David Mathews, Secretary of Health, Education and Welfare, United States Court of Appeals, Ninth Circuit, September 29, 1976.

18 *Dismukes v. Department of the Interior*, D.C. District Court, Dec. 19, 1984.

19 即Electronic Freedom of Information Act Amendments of 1996.

# 第二章　数据帝国的兴起

数据正在呈指数级增长。之所以增长速度如此之快，背后有许多原因。现在几乎所有数据的产生形式，都是数字化的。各种传感器的剧增，高清晰度的图像和视频，都是数据爆炸的原因。

如何收集、管理和分析数据正在日渐成为我们网络信息技术研究的重中之重。以机器学习[01]、数据挖掘为基础的高级数据分析技术，将促进从数据到知识的转化、从知识到行动的跨越。

联邦政府的每一个机构和部门，都需要制定一个应对“大数据”（Big Data）的战略。[02]

——《规划数字化的未来：美国总统科学技术顾问委员会给总统和国会的报告》2010年

如果说《信息自由法》在法律的层面上规定了政府机关的文件可以公开，其后的《阳光政府法》规定政府机关的会议必须公开，《电子信息自由法》又规定了计算机内的数据也不能例外，那么奥巴马继续开拓的空间似乎已经不大了。

他所谓的“要建设一个前所未有的开放政府”，究竟指的是什么呢？

奥巴马是哈佛大学法学院的高才生，他在就读期间，就担任了久负盛名的《哈佛法律评论》的主编；博士毕业之后，曾在律师行从业多年，还长期在哥伦比亚大学讲授《宪法学》。严谨的科班训练加上律师

实务生涯，他的逻辑思维是非常严密的。

他用“前所未有”（unprecedented）来形容他将要开创的事业，是因为，他清楚地知道：透明无止尽。对政府而言，只有更透明，没有最透明。虽然联邦政府的文件、会议甚至数据都规定了可以公开，但近年来信息技术突飞猛进的发展，特别是互联网的兴起，不仅给信息公开的内容、也给公开的方式带来了新的机遇和挑战。

这是个技术奔腾、信息爆炸的时代。奥巴马领导的联邦政府，正是美国社会的信息中枢。他的雄心，有广袤的用武空间。

# 摩尔定律：全世界半个世纪的发展规律

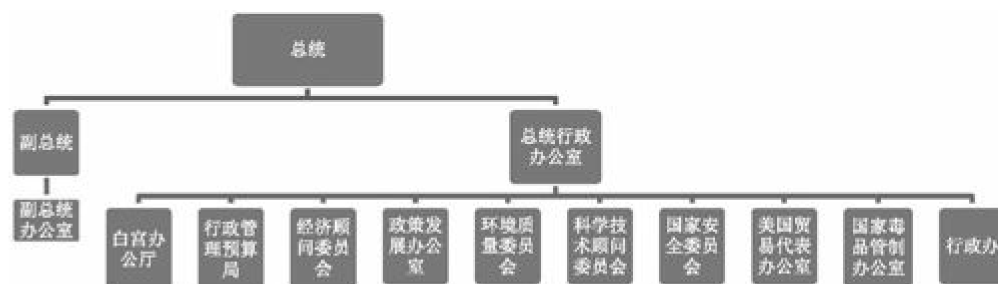
摩尔定律已经成为工业界一切呈指数型增长事物的代名词。……下一个十年，摩尔定律可能还将有效……可以肯定的是，创新无止境。[03]

——戈登·摩尔，英特尔公司创始人，2003年

联邦政府是美国最大的雇主，共雇用了约200万名工作人员。[04]

联邦政府主要由三大块组成：一是总统行政办公室（The Executive Office of the President），二是15个内阁部门（Cabinet Department），三是70多个独立的联邦机构（Independent Agency）。

总统行政办公室的结构




总统行政办公室是直接为总统服务的中枢部门。

其下辖白宫办公厅（The White House Office）、行政管理预算局（OMB）、经济顾问委员会（Council of Economic Advisers）、科学技术顾问委员会（PCAST）等机构。其中，最重要、最大的机构当属白宫办公厅和行政管理预算局，它们控制了信息、掌握了财权，是15个内阁部门和70多个独立机构的管理和协调单位，可谓中枢中的中枢。鉴于这

两个机构的重要性，本书将会多次提及。

作为全美最庞大的组织和机构，联邦政府也一直号称他们是美国最大的信息生产、收集、使用和发布的单位。[05]

数据和信息的区别

很多情况下，“数据”和“信息”两个词经常替换使用。但严格地说，数据和信息这两个概念有很大的区别：

数据是对信息数字化的记录，其本身并无意义；信息是指把数据放置到一定的背景下，对数字进行解释、赋予意义。

例如：“1.85”是个数据，“奥巴马身高1.85米”则是一则信息。

但进入信息时代之后，人们趋向把所有存储在计算机上的信息，无论是数字还是音乐、视频，都统称为数据。

如果要考察信息的多少，就必须以物理存储器上保存的数据量作为度量。因为所有的信息，都是以数据的形式保存在物理存储器上的。由于人类的数据量不断增多，近几十年来，科学家也相应定义了一些新的名词，来表示新的存储单位，以方便对客观世界的描述。

美国联邦政府到底收集了多少数据，其总量无从得知，但我们可以从现有的一些研究资料中窥见一斑。

理解数据的存储单位

| 单位 | 英语标识 | 大小 | 含意和例子 |
|---|---|---|---|
| 位 | Bit | 1 或 0 | 计算机用二进制存储和处理数据，一位是指一个二进制数位：0 或 1, 这是存储信息的逻辑单元。 |
| 字节 | Byte | 8 位 | 这是计算机存储信息的基本物理单位，存储一个英文字母在计算机上，其大小就是一个字节。 |
| 千字节 | KB | 1024 字节，或 $2^{10}$ 个字节 | 一页纸上的文字大概是 5 千字节 |
| 兆字节 | MB | 1024 千字节，或 $2^{20}$ 个字节 | 一首普通 MP3 格式的流行歌曲大概是 4 兆字节 |
| 吉字节 | GB | 1024 兆字节，或 $2^{30}$ 个字节 | 一部电影大概是 1 吉字节 |
| 太字节 | TB | 1024 吉字节，或 $2^{40}$ 个字节 | 美国国会图书馆所有登记的印刷版书本的信息量为 15 太字节，截至 2011 年年底，其网络备份的数据量为 280 太字节，这个数据还在以每个月 5 太的速度递增。 |
| 拍字节 | PB | 1024 太字节，或 $2^{50}$ 个字节 | 美国邮政局一年处理的信件大约为 5 拍，谷歌每小时处理的数据为 1 拍 |
| 艾字节 | EB | 1024 拍字节，或 $2^{60}$ 个字节 | 相当于 13 亿中国人人手一本 500 页的书加起来的信息量 |
| 泽字节 | ZB | 1024 艾字节，或 $2^{70}$ 个字节 | 截至 2010 年，人类拥有的信息总量大概是 1.2 泽字节 |
| 尧字节 | YB | 1024 泽字节，或 $2^{80}$ 个字节 | 超出想象，难以描述 |

（部分例子参照了All too much，The Economist，2010年2月25日）

**2009**年美国各行业数据存储量对比




（数据来源：International Data Corporation）

2011年5月，麦肯锡公司下属的全球研究所（McKinsey Global Institute）出版了一份专门的研究报告《大数据：下一个创新、竞争和生产率的前沿》。[06]该报告对美国政府目前拥有的数据量进行了估算，在制造业、新闻业、银行业、零售业等17个行业当中，美国政府共拥有848拍字节（Petabyte）的数据总量，仅次于离散式制造业[07]的966拍，居第二位；居第三位的是新闻传媒业，共有715拍字节。

这是美国政府作为一个行业的总体情况，下面我们来考察联邦政府中具体的单个组织。

以商务部下属的美国普查局（USCB）为例，它目前拥有2560太字节（Terabyte）的数据。“太”，代表2的40次方，它的大小，已经大大超出了人类的直接感知能力，只能通过形象的比喻来描述：如果把这些数据全部打印出来，用4个门的文件柜来装，需要5000万个才能装得下。沃尔玛是世界上最大的零售王国，它每小时要处理100多万笔电子交易记录，可谓每分每秒都在源源不断地生产数据；2010年，其数据库大小为2500太字节左右，还没有赶上美国普查局。

除了美国普查局，国家安全局（NSA）和中央情报局（CIA）都拥有超级巨大的数据库。2011年5月，历经十年，美国人终于在巴基斯坦将本·拉登击毙，报了“9·11”的一箭之仇。帕拉契尼（John Parachini）是兰德（Rand）公司情报政策研究中心的主任，他在接受《巴尔的摩太阳报》的采访时介绍说，国家安全局是从电话监控的记录当中发现了本·拉登的蛛丝马迹。[08]该局对全美的电话进行监控，所收集的数据量是惊人的，它每6小时产生的数据量就相当于美国国会图书馆所有印刷体藏书的信息总量。而美国国会图书馆，是世界上馆藏量最大的图书馆。

再说中情局，其本职工作就是收集情报信息。业内专家普遍认为，其数据库比普查局、国安局的还要大，很可能拥有全世界最大的数据库。

普查局、国安局、中情局只是联邦政府数百个机构当中的几个例子，还有财政部、卫生部、劳工部，这些都是数据密集型的行政管理部门。即以财政部为例，根据行政管理预算局的信息收集年度报告，2009年，财政部因为收集信息产生的社会负担为76亿小时，占全部联邦政府收集信息社会负担总数的78％，[09]之所以如此，是因为收税和退税的过程极为繁琐，但76亿小时收集工作会产生多少数据量，其大小也难以想象。

再换一个角度，我们来看看这个联邦政府的硬件资产。

1998年，联邦政府共拥有432所数据中心，专门负责各类数据的存储和维护工作。2010年，数据中心的总数跃升到2094所，翻了几倍。

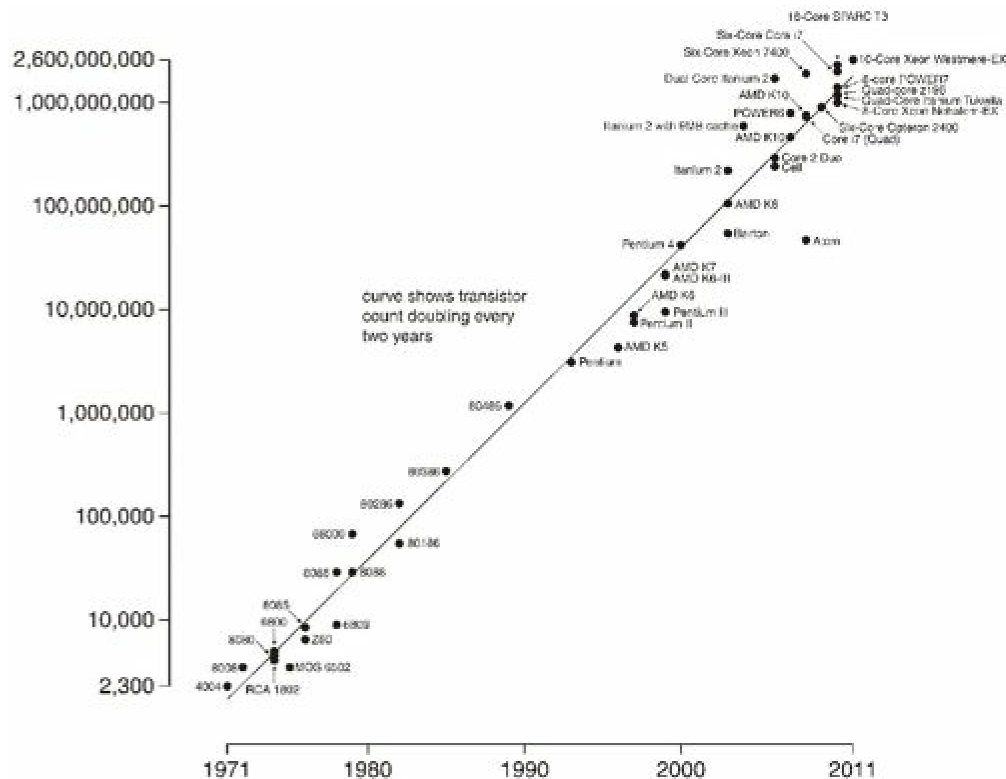
庞大的数据资产，是需要经费来支持的。1996年，联邦政府的年度IT预算是180亿美元，十多年来不断地上升，2010年，已经高达784亿美元；由于连年巨额的投资，联邦政府已经声称，他们是全世界范围之内最大的信息技术消费者。而据报道，这些投资中的一半以上，都用在了购买存储数据的硬件设备上。

这是一个不折不扣的数据帝国。

帝国形成的原因，已经有很好的解释，这就是摩尔定律（Moore's Law）。

1965年，英特尔（Intel）的创始人之一戈登·摩尔（Gordon Moore）考察了计算机硬件的发展规律，提出了著名的摩尔定律。该定律认为，同一个面积集成电路上可容纳的晶体管数目，一到两年将增加一倍，[10] 也就是说，其性能将提升一倍。换句话说，计算机硬件的处理速度和存储能力，一到两年将提升一倍。

**1971**年至**2011**年不同中央处理器上的晶体管数量和摩尔定律




说明：纵坐标为晶体管数量，横坐标为年份。图中圆点表示不同品牌的中央处理器。该曲线表明，从1971年至2011年，大概每两年同一面积大小中央处理器集成电路上的晶体管就增加了一倍。（图表来源：维基百科）

回顾这近半个世纪的历史，硬件技术的发展，基本符合摩尔定律。摩尔定律的一个重要结果，是推动了全世界对物理存储器的消费；其消费量增加的速度，有学者认为，甚至比摩尔总结的硬件发展速度还要快，从1990年代起，全世界的物理存储器，每9个月就增加一倍。[11]



戈登·摩尔

出生于1929年，至今健在，他于1956获得加州理工学院的博士，1968年成为英特尔的创始人之一，也是摩尔定律的首创人。（图片来源：英特尔公司网页）

其中的原因，是因为物理存储器的性能不断提高，同时价格还不断下降。1955年，IBM推出了第一款商用硬盘存储器，每兆字节的存储量需要6000多美元。此后，硬盘存储器的价格以越来越大的加速度下降。1993年，购买1兆字节的存储量只需大概1美元；2010年，这个价格下降到不足1美分。多数专家都相信，计算机硬件的技术将持续发展，价格还将下跌，直到2020年，摩尔定律还将仍然有效。

计算机硬件这种令人“瞠目结舌”的发展速度，使全世界的数据处理

和存储不仅越来越快、越来越方便，还越来越便宜，海量数据的积累最终成为可能。

但像盛水的杯子一样，存储器毕竟只是容器。关键的问题在于，帝国的数据从何而来？海量数据的源头在哪里？

美国联邦政府的数据来源，当然首先缘于它各个部门的业务工作，也就是业务数据。

作为社会管理和公共服务的提供部门，收集数据、使用数据，是自古以来全世界的政府都在普遍采用的做法。但政府开始大规模、系统地收集数据，其历史并不久远。在美国联邦政府的发展历史上，业务数据的收集，有一个重要的里程碑，这就是“最小数据集”。

# 最小数据集：上升到立法高度的开路先锋

一个好的数据结构和一个糟糕的代码，比一个糟糕的数据结构和好的代码要强多了。[12]

——埃里克·雷蒙

美国软件开源运动的领导者，1999年

最小数据集的概念起源于美国的医疗领域。

1973年，在国家生命健康统计委员会（NCVHS）的主导下，为了规范出院病人的信息收集工作，美国第一次制定了统一的出院病人最小数据集，既然是出院，核心的环节就是付钱，所以这些数据不久后又被用于创建统一的医疗账单（Uniform Bill，UB）。

1975年，美国医院协会（AHA）成立了国家统一账单委员会。经过了几年的讨论，1982年，该委员会出台了UB-82的数据格式，统一了全国的医疗账单格式。1992年，UB-82又被修改升级到UB-92，并被扩大应用到了医疗保险和索赔的领域。

由于其实用性，最小数据集的概念在医疗领域被迅速推广。近几十年以来，几乎每年都有新的最小数据集被定义、开发和推广。目前，已经被应用到眼科、牙科、皮肤科、妇科以及体检、护理、急救、住院等医疗服务的方方面面，衍生出各种各样的、特定的最小数据集。

随着时间的推移，“最小数据集”在美国已经演变成了一个一般性的概念，它指代国家的管理层面针对某个业务管理领域强制收集的数据指

标。不少领域的“最小数据集”甚至被上升到立法的高度。例如，对于养老院的管理，美国国会就规定，每个养老院都必须提交一系列关于老人健康指标的最小数据集给州政府的医疗管理部门，该部门汇总之后，再提交联邦政府的管理部门。

当然，何为“最小”，政府的管理者、决策者和公共服务的提供方、接受方都有不同的需求和看法，很难达成一致。某一特定的数据指标是否应该纳入，不同的立场、视角、环境和管理水平都会导致不同的意见。而且，各个最小数据集本身可能是完整的、有效的，但当各行各业的最小数据集越来越多之后，从全局出发，它却不一定是合理的，因为可能存在更好的划分方法，使各个最小数据集之间具有更明确的边界和更少的重叠。

最小数据集（Minimum Data Set, MDS）

最小数据集是指通过收集最少的数据，最好地掌握一个研究对象所具有的特点或一件事情、一份工作所处的状态，其核心是针对被观察的对象建立一套精简实用的数据指标。

因此，每一个“最小数据集”的出台，都意味着多年的纷争和详尽的论证。

最小数据集的出现，最早是因为不同组织之间信息交换的需要，例如，两个医院之间，医院和政府医疗管理部门，医院和保险公司之间以及一些社会福利部门之间，都有交换信息的需要。随着最小数据集的推广，越来越多的社会组织、地方政府和联邦政府的业务部门之间都建立了标准的“数据接口”，从此彼此“数据”相连。

但到了1980年代，一场新的技术浪潮又把最小数据集的应用推上了新的高点。

这就是信息管理系统的兴起。

1975年，比尔·盖茨创办了微软。次年，史蒂夫·乔布斯成立了苹果电脑公司。之后，个人电脑、商业软件开始得到大面积普及，开发新的信息管理系统开始成为各行各业迈向信息化的主要措施。

所谓的“信息管理系统”，也就是实现某一特定业务管理功能的软件。

软件的构成，主要有两部分，一是程序（也可称为代码），二是数据（或称为数据库）。程序和数据的关系，就好像发动机和燃料，所有的程序，都是靠数据驱动的；数据之于程序，又好比血液之于人体，一旦血液停止流动，人就失去了生命，代码也将停止运行。

数据的生命力，甚至比程序更持久。程序可以不停地升级、换代甚至退出使用，但保存数据的数据库却会继续存在，其价值很可能与日俱增、历久弥新。世界万维网之父蒂姆·伯纳斯—李，曾经在2006年这样论述说：

“数据是宝贵的，它的生命力，比收集它的软件系统还要持久。”[13]

对于软件开发而言，数据库的设计甚至比程序的设计还要重要。埃里克·雷蒙，是美国软件开源运动的领袖，他在谈到代码和数据时曾表示：

“一个好的数据结构和一个糟糕的代码，比一个糟糕的数据结构和好的代码要强多了。”

最小数据集，其实就是一个业务管理过程当中最重要的数据指标。它在各个公共领域的定义和推广，成了这些部门在开发设计信息管理系统时最重要的一个参考，因为一旦核心的数据收集指标被确定，数据库的结构设计就成为一个水到渠成的过程。有些最小数据集，甚至直接就被引用，成为信息管理系统的数据结构。

就好像开路先锋，最小数据集为信息管理系统的开发和设计起了重要的铺垫作用。1970年代以来，随着计算机的普及，美国产生了越来越多的最小数据集，各种信息管理系统也开始大幅增加。截至2011年，美国的联邦政府已经拥有1万多个独立的信息管理系统。[14]

几乎每一项业务，每一个新的立法、新的计划，都会有一个数据库和信息管理系统与之对应。因为，没有任何一项工作，不涉及收集信息，而这些都离不开数据。

如今，联邦政府可谓事无巨细，都有一个信息系统在管理，其背后的数据库可谓五花八门，多不胜举。

例如，美国现在债台高筑、不断冲击上限，联邦政府甚至一度产生财务危机，奥巴马也为此头痛不已。联邦政府财政部下属的公共债务局（BPD）是国家债务的主管部门，要追踪这样一笔庞大国债的来龙去脉，该局自然拥有不少信息管理系统，其中，有一个是专门用来记录“捐款”的。

“欠债”和“捐款”，听起来似乎风马牛不相及。但在美国，两者还是拉上了关系，这是因为有些美国人试图通过自己的捐助来缓解国家的财政负担。公共债务局就为此专门建了一个信息系统，来记录美国公民为减少国债作出的个人捐赠。数据表明，2010财政年度，该局共收到

2840466.75美元的捐款，2009年的捐款曾突破300万元，为3063057.05美元。捐款数量的下降，表明了人们的可支配收入在减少，也从另外一个侧面证明了美国经济确实在衰退。

历史上最大的一笔个人捐赠来自于2006年，一位俄亥俄州的老人在去世之后将自己价值110万美元的财产捐赠给了联邦政府公共债务局。当然，这些捐款，对于美国14.6万亿的债务来说，只是杯水车薪。但系统的管理，不仅笔笔在案、账目清晰，便于统计分析，也体现了对捐赠者的尊重。

# 民意几时有：选票催生的创新

我想要成就的事情，就是我的人民想要做的事情；我的任务，就是要准确地发现人民的需要。[15]

——亚伯拉罕·林肯，第16任美国总统

一般来说，业务数据都由下级部门和各类社会组织通过“数据接口”上报给联邦政府。

但作为一个中央政府，只接受数据是远远不够的，联邦政府也需要走出去，主动收集数据，了解全社会对某项政策的评价、单个公民对某个问题的看法，这就是所谓的民意调查、社会调查。

> 民意调查（Public Opinion Poll）
>
> 民意调查是指通过对一小群、有代表性人口的调查和访谈，预测社会全体公众对一些政治、经济和社会问题的态度和看法。其本质是“观一斑而知全豹”、“观一叶而知秋”。

美国的各种调查（Survey）之多，可谓铺天盖地，这些调查的直接目的，就是收集、掌握、分析反映民意和社情的第一手数据。




1948年，乔治·盖洛普登上《时代》杂志的封面。

这一年，杜威和杜鲁门竞选总统，盖洛普预测杜威将胜出。开票结果出来的当晚，《纽约时报》、《生活》、《芝加哥论坛报》等报刊都印刷好了杜威的照片，并题为“美国新总统”，但最终杜鲁门胜出。这次预测失败，又引起了人们对民意调查的质疑。盖洛普总结原

因说：因为两人的民调一直相差太大，他们在大选前3周就提前结束了调查。此后，盖洛普持续改进调查的方法，最终在美国乃至在全世界都赢得了巨大的声誉。现在，盖洛普的品牌已经成为民意调查的代名词。

美国的民意调查最早源于对总统大选投票结果的预测。

1824年，位于宾夕法尼亚州的一家报纸*Harrisburg Pennsylvanian*第一次发布了关于谁能当选总统的预测。虽然其预测最后被证明是错误的，但却被大众一再津津乐道。

此后，各大报纸杂志都不想错过这个“抢眼球、聚人气”的话题，争相开展民意调查，以期准确地预测到底谁能当选总统，这开启了民意调查的时代。

蜂拥而上的结果，就是竞争。竞争的结果，就是民意调查的科学性不断提高，范围不断扩大；在20世纪30年代的美国，最终形成了一个社会调查的产业。

其中的转折点，是1936年。

这一年，第32任美国总统富兰克林·罗斯福为了争取连任，与共和党的兰登（Alfred Landon）对垒，打响了选战。

这时候，一本叫做《文学文摘》（*Literary Digest*）的杂志风头正劲。

《文学文摘》成立于1890年，其畅销的主要原因，是因为它准确地预测了1920、1924、1928、1932年等4届总统大选的结果，随着该杂志

销售量的年年攀升，民意调查的热度和可信度也不断上升。

1936年，《文学文摘》在对240万普通公众进行了调查之后，把“宝”压在了兰登的身上。这个时候，一家刚刚成立不久的研究所，只对5000人进行了调查，却宣布罗斯福会胜出。

这家研究所就是1935年成立的美国舆论研究所（AIPO），它的奠基人，是美国民意调查科学化的先驱：乔治·盖洛普（George Gallup）。

罗斯福最终以大比分击败兰登，成功连任；盖洛普也取代了《文学文摘》，成了新的行业领袖。这一仗，成了《文学文摘》的“滑铁卢”，该杂志次年就宣布破产、退出市场。5000人的问卷击败了240万人的调查，盖洛普领导的美国舆论研究所当然随之身价倍增、名扬全国。

盖洛普的成功，根本原因在于他掌握了一套科学的人群抽样方法，而不是盲目的大面积访谈。此后，一大批新的、专业化的民意调查机构应运而生，调查方法的科学性不断增强。从1936年到2008年，共举行了18次总统选举，盖洛普民调（Gallup）成功地预测了16次。[16]

报纸举办民意调查，其目的是为了一“鸣”惊人，制造新闻效应，扩大报纸的声誉和销量。但对总统候选人来说，他们也要参考民意调查的结果，因为他们对民意的掌握，往往关系到选举的最终成败。

这是因为，在民主社会，谁上谁下，人民的选票有最终的话语权。

票多，则胜。

作为候选人，争取选票的唯一方法就是争取民意。要争取民意，首先就要了解民意。现代政治学中，有一个“中间选民理论”很好地解释了

其中的奥秘。

中间选民理论（Median Voter Theory）

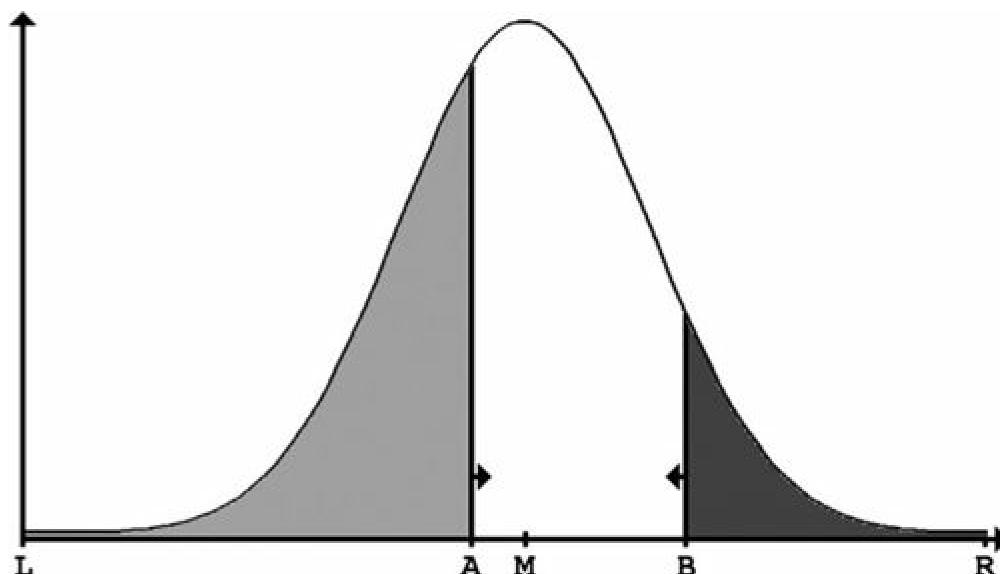
该理论也是在1940年代提出的。在选举中，所有的选民都有自己的个人偏好，这个偏好对应于平面坐标上的一点，全部选民的偏好将呈正态分布，也就是一个钟形曲线，这个曲线就是民意分布图。

每个选民最终都会把自己的选票投给与自己意见最接近的候选人。在多个候选人竞争的情况下，候选人想要胜出，就要争取最多的选票，他就必须找准民意最集中的地方，为最多的民众说话，这一点，就是民意的“中值”。

中间选民理论提出，民意有一个“中值”，而民意调查，正是帮助候选人找准民意“中值”的“阿拉丁神灯”。

罗斯福虽然在当选总统前就半身瘫痪，长期坐在轮椅上，但后世却常常把他和华盛顿、林肯相提并论，公认为美国历史上最伟大的总统之一。作为唯一一位连任过4届的总统，他深谙民意调查的作用和潜力。他不仅在竞选中应用民意调查，1940年，他还正式将民意调查的方法引进到了联邦政府的政策制定过程当中，成为美国历史上第一位将民意调查和公共政策相结合的总统。这些成就，使他成为美国历届总统当中至今无人超越的“民意大师”。

民意的“中值”和选票多少的关系




说明：图中的钟形曲线即代表民意的分布，左边浅色和右边深色的区域面积大小，分别代表候选人A和候选人B可以预期得到的选票多少。M代表民意的中值点，候选人A和B，谁能确定M点，不断向M点靠拢，谁就能获得更多的选票。

上图表明：候选人A比候选人B更接近民意的中值，将获得更多的选票。

罗斯福之后，几乎每一位美国总统，不论是竞选还是执政，都大规模地采用民意调查的方法。

随着计算机技术的不断发展、统计分析水平的提高，民意调查逐渐成了联邦政府了解社会、掌握民情最有效的工具和方法。调查的范围也不断扩大，从普通公民的观点和看法，慢慢延伸到个人的行为、事实和社会现象，最后发展成了更广泛意义上的社会调查。

到今天，可以说，联邦政府的各个职能部门都在开展社会调查。其中，有两个机构堪称联邦政府社会调查的“航空母舰”，它们每年开展无

数大大小小的社会调查，其他机构都望尘莫及。这两艘“航空母舰”，一是劳工部下辖的劳工统计局（BLS），一是商务部下属的普查局（BC）。



富兰克林·罗斯福（1882-1945）

罗斯福从1921年起下半身全部瘫痪，但这位残疾人却在1933年至1945年连续4次当选总统，带领美国人走过了第二次世界大战的艰难时段。对于民意，他强调：不仅要掌握民意，还要领先民意一小步，才是民主政治中的胜选之道。（图片来源：维基百科）

劳工统计局对就业市场开展很多专业调查，其中最大的一种被称为“国家纵向调查”（National Longitudinal Survey）。所谓“纵向”，是指以时间轴为单位，在确定调查对象之后，对其进行长期的跟踪、反复的问卷，收集大量的数据，然后进行统计分析。

1966年，劳工统计局开展了美国历史上第一个国家纵向调查。调查对象是全国的男性，分为青年和老年两组，该局在全国范围内锁定了10245名男性，进行了长达24年的全方位跟踪调查，直到部分老人撒手归西为止。类似的国家纵向调查项目一共有4个。1997年最新铺开的一个，调查对象是1980年至1984年出生的青少年，简称为NLSY97（National Longitudinal Survey of Youth 1997），该局在全国范围内确定了约9000名青少年，每年都进行一次特定话题的跟踪问卷调查，至今还没有结束。

以1997年的青少年纵向调查为例，它跟踪调查的对象虽然是青少年本身，但受访人、问卷人却可以是家长、老师、雇主等等和该青少年密切相关的个人和群体，每次问卷都有数十个问题，需要受访人1小时左右才能完成。因为付出了时间，受访人在完成问卷之后，能获得8到20美元不等的报酬。同样的问题，也可能在不同的年份、在相同受访人员的问卷中重复出现，以测试受访人态度的变化。

NLSY97虽然是由劳工统计局负责实施，但却是一个综合项目，涉及青少年生活的方方面面，该局邀请了多个政府部门参与问卷的设计。例如，青少年犯罪预防办公室（OJJDP）隶属于美国司法部，是青少年犯罪问题的国家主管部门，他们帮助劳工统计局设计了有关青少年犯罪方面的问卷。就青少年犯罪而言，问题可能包括轻度违法记录、毒品滥用情况、酗酒情况、性行为、家庭成员的构成及背景、收入、教育、家庭关系、和父母的关系、青春期情况、约会情况、恋爱情况、生育情况、职业培训情况、接受社会福利项目的援助情况、个人时间管理、生活目标的变化、学习成绩、宗教活动、所住社区的人口特点、所在城市的经济特点，等等。

随着数据的积累，一幅以个人成长为中心、越来越大的社会画卷也

开始展开。这种以一个国家为单位的大型社会调查，是研究一个社会长期变迁不可或缺的重要资源，也为政策的制定、调整和评价提供了重要的参考和依据。

另一艘“航母”是普查局。和劳工统计局相比，它的社会调查名目更多、范围更广。除了负责十年一次的人口普查之外，该局每年还在全国范围之内开展系统的行业调查，其中最著名的有：全国社区调查、全国消费者开支调查、全国医院调查、全国建筑调查、全国房屋调查，等等。

除了掌握民情、了解社会，问卷调查还是联邦政府评估资金使用绩效的主要工具。联邦政府每年都要下拨大量的专项资金，资金的使用效果和好坏，往往是来年是否继续拨款的决定性因素，联邦政府目前最重要的衡量手段，也是社会调查。

以卫生部下属的药物滥用和精神健康管理局（SAMHSA）为例，每年，该局都有专项基金，用于社区的毒品防治工作，全国的基层政府、公益组织都可以申请。但这笔钱也很“烫手”，拿了之后，有大量的问卷调查工作要开展、要上报。

社区的毒品防治工作主要是政策宣讲和培训。申领了专项资金的单位，就要开展培训，每个参加培训的人都要参加4次问卷调查，分别在培训开始前一次、培训中间一次、培训结束时一次，三个月以后还有一次。这些问卷调查的数据，基本都是通过互联网汇总上报。该局通过对比分析4次调查结果的具体指标，评估该单位的资金使用效果，再决定下期资金是否下拨。

联邦政府的这些调查，虽然目的不同、性质不同、方式不同，但最

后产生的结果，无一例外，都同样是数据。

# 普适计算：计算机本身将从人们的视线中消失

最高深的技术是那些令人无法觉察的技术，这些技术不停地把它们自己编织进日常生活，直到你无从发现为止。[17]

——马克·韦泽，普适计算之父，1991年

1988年，互联网的概念刚刚兴起。这种新的信息共享方式，令全世界都兴奋不已。当时，绝大部分科学家，都还在品味和研究网络带来的巨大便利，沉浸在欣喜当中。

这时候，美国施乐公司（Xerox）的一名计算机科学家，却开创性地提出了“普适计算”（Ubiquitous Computing）的理论，为网络在人类未来生活中的作用以及计算方式的改变作出了前瞻性的预测。

他就是马克·韦泽（Mark Weiser）。

韦泽是密歇根大学毕业的计算机博士，青年时代就表现出杰出的天才。他毕业之后，曾经在马里兰大学任教8年，其后加入了施乐公司的帕罗奥多研究中心（PARC）。

帕罗奥多研究中心是全世界著名的创新中心，鼠标、激光打印机、以太网、语音压缩技术等等伟大的发明都在这里诞生。

韦泽领导了帕罗奥多研究中心计算机科学实验室的发展，1996年又担任了该中心的首席技术官。

韦泽认为，自从计算机发明以后，人类的计算方式，先后经历了两

个阶段：一是主机型计算阶段（Mainframe Computing），指的是很多人共享一台大型机；二是个人型计算阶段（Personal Computing），指的是每一个人都可以拥有一台电脑。韦泽预测，由于网络技术的兴起，特别是无线网络技术的发展，计算机本身将从人们的视线中消失，计算将最终和环境融为一体。人们能够在任何时间和任何地点获取、处理信息，这就是普适计算的阶段，人类正在向这第三波计算浪潮迈进。

普适计算

通过在日常环境中广泛部署微小的计算设备，人们能够在任何时间和任何地点获取并处理信息，计算将最终和环境融为一体。这就是普适计算，是人类的第三波计算浪潮。

一句话：万事万物，凡存在，皆联网，凡联网，皆计算。



马克·韦泽（1952-1999）

曾为帕罗奥多研究中心首席技术官，被称为普适计算之父。

韦泽还是一名摇滚乐队的鼓手，在他的策划下，其乐队在1993年实现了美国互联网历史上的首次现场直播。他因此得名“摇滚乐队当中最聪明的鼓手”。（图片来源：维基百科）

实现普适计算的根本，是在人类生活的物理环境当中广泛部署微小的计算设备。

无处不在的微小计算设备和无处不在的互联网相结合，实现无处不在的信息自动采集、传递和计算。

这种微小的计算设备，就是传感器。近年来流行的物联网概念就是普适计算的最佳例子。

对于传感器及其网络的最早研究，始于美国国防部一个军事项目的研究，后来技术日臻成熟，传感器的应用逐渐从军事领域扩大到民用领域。

这可以追溯到20世纪60年代。

1962年，一场代号为“圣灰星期三”（Ash Wednesday）的风暴席卷了美国东海岸600多英里的海岸线，这场风暴持续了3天，影响了全美6个州，最后造成了40人死亡、1000多人受伤，导致了几亿美元的经济损失，被后人评为20世纪美国最严重的十大风暴之一。

由于损失惨重，美国国会对救灾防灾工作召开了专门的听证会，最

后促成了军民联手的“海浪监测计划”：美国陆军工程部和美国国家海洋与大气管理局（NOAA）共同建设一个传感器监测系统，对兴风作浪的海洋进行监测。

这项计划的实施结果，是在全美海岸线和五大湖区建立了一个定点的、连续的、实时的传感器网络，对海浪的大小进行监控。受限于当时的技术，最早的传感器只能监测海浪的能量。从2005年起，美国国家海洋与大气管理局在浮标上装备了更高端的传感器，开始监测海浪的方向。

2009年，系统再次升级。该局开始着手建立一个覆盖全美海岸线、从浅水到深水的、精确的海浪监测网络。这个网络总共在近海、外大陆架、内大陆架和沿海设置了296个传感器。新的传感器不仅能监测海浪的能量和方向，还能计算它的传播速度、偏度和峰度。[18]

这些传感器以分秒为单位，将数据源源不断地实时传回到国家海洋数据中心（DODC）。

对海浪的监测，不仅能提高沿海地区对海啸、风暴等自然灾害的应急能力，还能极大地改善海上的交通安全。根据美国疾病防控中心（CDC）的统计，捕鱼业是美国最危险的职业之一，全美所有行业的平均致死率为0.004％，而捕鱼业的平均致死率高达0.155％，其中79％的死亡是天气变化的原因导致的。[19]

除了安全，海浪的监测还能为利用大海能量进行发电提供关键的分析型数据。

海浪监测只是联邦政府利用传感器网络自动采集数据、迈向普适计算的一个例子。事实上，由于无线传感器的快速发展，普适计算已经在

美国的农业、运输、能源和建筑等领域逐步铺开。

2011年10月，联邦政府商务部下属的国家气象局（NWS）宣布，该局已经在全国2000辆客运大巴上装备了传感器，随着巴士的移动，这些传感器可以收集沿途所有地点的温度、湿度、露水、光照度等数据，并立刻传回国家气象局的数据中心。数据采集是每10秒钟一次，每天传感器要采集10万次以上的数据。这些数据是实时的、高精度的，这意味着，天气预报将不再仅仅是“预”报，将逐渐走向“实”报、“精”报。

此外，联邦政府国家邮政局（USPS）也宣布，他们正规划在全部邮车上安装传感器，在邮车投递邮件的同时，实时采集社区的空气质量、污染指数和噪声等数据指标。

有评论家感叹道：谁也没想到，汽车，这个工业时代的标志和先锋，如今又成为信息时代普适计算的“排头兵”。

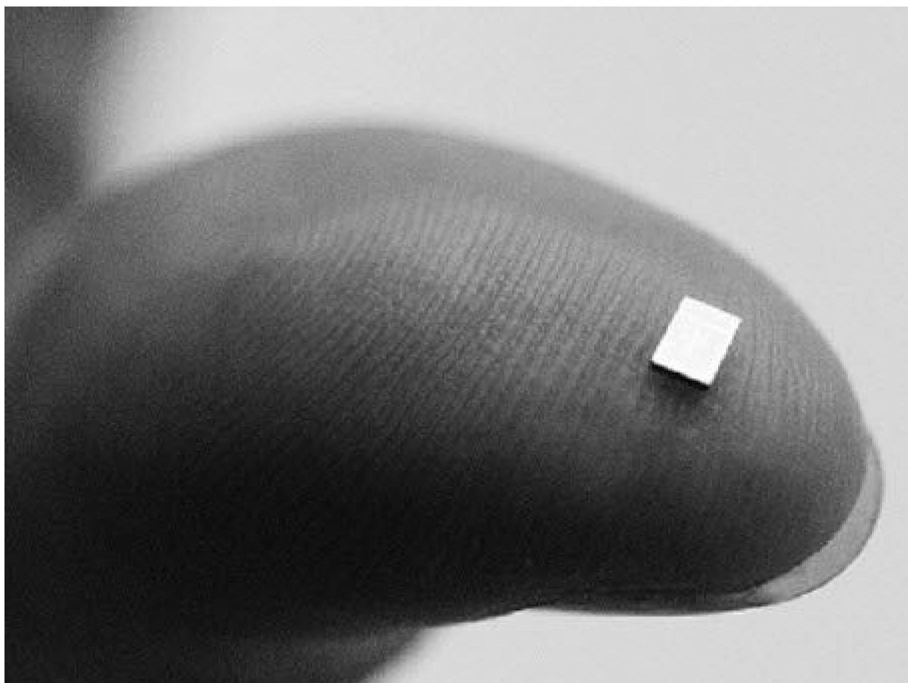
近年来，传感器的发展可谓突飞猛进。一种新的无线传感器：射频识别标签（Radio Frequency Identification，RFID），正异军突起，也在美国联邦政府得到了大规模的应用。



薄如纸张的RFID




RFID动物耳标



更小更薄的RFID

（图片来源：网络）

RFID精巧轻便，既可以薄如纸张，也可以小如豆粒，却能无线存储、发送、读写数据，目前的应用主要集中在身份标识领域。以农牧业为例，1990年以来，全球各地陆续爆发动物疫情，2003年12月，美国发现了第一宗疯牛病病例。2004年起，联邦政府农业部启动了“全国动物身份识别系统”（National Animal Identification System）的项目，为全国的新生牲畜建档立户、配置射频识别耳标。通过这个移动传感器，对牲畜进行连续跟踪，一旦家畜疫情爆发，就能通过数据库追踪溯源，快速确定传染源和传播范围。美国现在已经装备射频识别耳标的家畜总数，无从得知，但可以肯定，这个数据库，也是海量级的。

从2005年起，美国食品与药品管理局（FDA）开始在药品上推行配备RFID的做法，以打击假药。美国国务院也开始颁发带有RFID标识的护照，以打击假护照，方便出入境的管理。

美国联邦政府通过传感器自动采集数据的例子，正在大幅增加。如果仅仅从数据量上来看，通过传感器自动采集的数据，已经取代了人工收集的业务数据，成为其最大的数据来源。还可以肯定的是，随着人类向普适计算不断迈进，通过传感器自动采集的数据将持续“爆炸”。

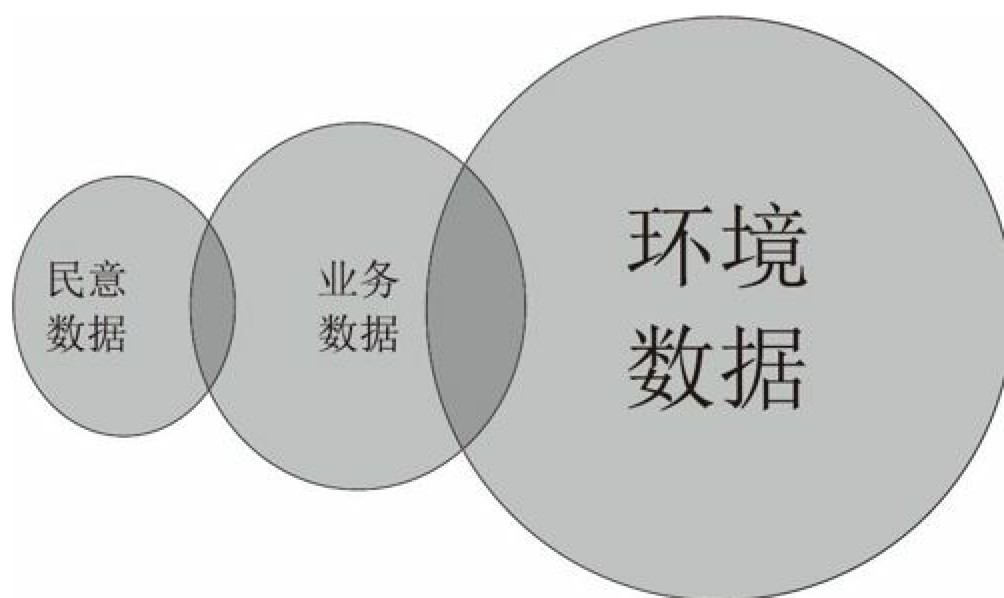
## “大数据”战略：争夺全世界的下一个前沿

联邦政府的每一个机构和部门，都需要制定一个应对“大数据”的战略。[20]

——《规划数字化的未来：美国总统科学技术顾问委员会给总统和国会的报告》2010年

业务工作的管理数据，民意社情的调查数据，以及对大自然、动植物的特点和变化进行监控而产生的环境数据，是联邦政府的三大数据来源。这三种数据，其发展各有先后，收集方式各不相同，数据量也大小不一。它们之间，存在着一些交叉和重叠，有一些民意调查的数据，是业务数据，而一些因环境监控产生的数据，也可以是业务工作的数据。

联邦政府三种数据源的关系和数据量的大小比较




联邦政府三种数据源以及收集方式的对比

| 数据种类 | 收集对象 | 收集方法 | 里程碑 |
| --- | --- | --- | --- |
| 业务数据 | 下级部门和各类社会组织 | 以基层上报、被动接收为主 | 1973 年卫生部引进最小数据集 |
| 民意数据 | 单个公民或组织 | 投入人力财力、主动收集 | 1940 罗斯福总统引进民意调查 |
| 环境数据 | 自然环境、动植物、物体 | 以传感器自动采集为主 | 1962 年的海浪监测计划 |

由于无线传感器的快速普及，环境数据增长得最快，成为联邦政府数据量最大的来源。

虽然环境数据增长得最快，但这三种数据，其实都在爆炸。这种爆炸，并不仅仅是数量一个维度的。2001年，著名的高德纳咨询公司（Gartner）在一份研究报告[21]中指出，数据的爆炸是“三维的”、是立体的，这三个维度，主要表现在以下三个方面：

一是同一类型的数据量在快速增大；

二是数据增长的速度在加快；

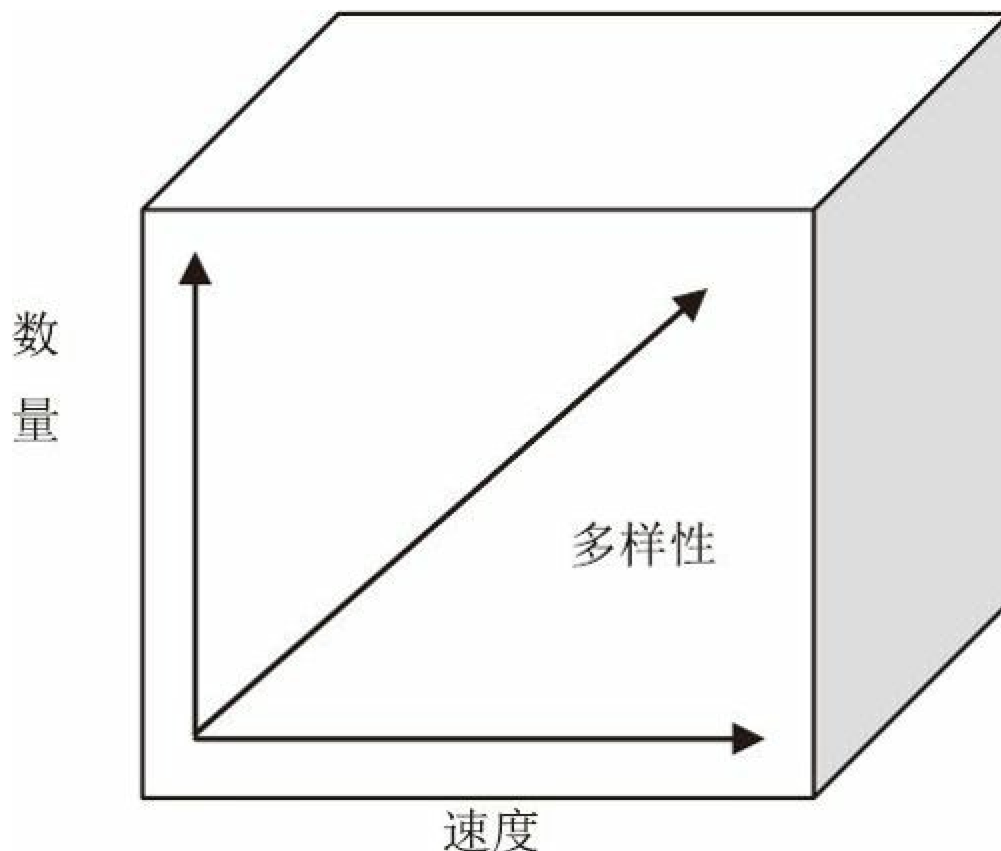
三是数据的多样性，即新的数据来源和新的数据种类在不断增加。

数据的爆炸性增长，也不仅仅限于联邦政府。如前文所述，2011年麦肯锡公司在其研究报告《大数据：下一个创新、竞争和生产率的前沿》中指出，在美国，仅仅制造行业就拥有比美国政府还多一倍的数据，此外，新闻业、银行业、医疗业、投资业、零售业都拥有可以和美国政府相提并论的海量数据。

哈尔·范里安（Hal Varian）是谷歌的首席经济学家，也是美国研究信息经济学的著名学者。2000年，他对数据和信息产生的速度进行了研究，他认为，人类社会每年产生的信息量，实在太大了，已经没办法用准确的方法来计算现有的数字信息总量，只能估算。他估计2000年新产生的数据量为1000拍到2000拍。但到2010年，仅仅全球企业一年新存储

的数据量就超过了7000拍，而全球消费者新存储的数据量约为6000拍。

数据的三维增长




这种数据量的增长，已经大大超出了人类的预期和想象。时至今日，数据已经像“洪流”一样，在全球的政治、经济生活当中奔腾。而且，随着信息技术的普及和进步，新的支流还在不断产生，各个支流流动、交汇和整合的速度，还在继续加快。

作为美国社会的信息枢纽，联邦政府当然要正面迎对这个挑战。

2010年12月，总统行政办公室下属的科学技术顾问委员会（PCAST）、信息技术顾问委员会（PITAC）向奥巴马和国会提交了《规划数字化未来》的专门报告，该报告把数据收集和使用的工作，提

到了战略的高度。

这个报告列举了5个贯穿各个科技领域的共同挑战，报告指出“每一个挑战都至关重要”，而第一个挑战就是“数据”问题。报告说：

“如何收集、保存、维护、管理、分析、共享正在呈指数级增长的数据是我们必须面对的一个重要挑战。从网络摄像头、博客、天文望远镜到超级计算机的仿真，来自于不同渠道的数据以不同的形式如潮水一般向我们涌来。这些数据以不同的格式存储在不同的环境中，有的在计算机的硬盘里，有的在数据仓库之内。

如何保证这些数据现在、将来的完整性和可用性，我们面临着很多的问题和挑战。如何使用这些数据，则是另外一个挑战……应对好这些挑战，将引导我们在科研、医疗、商业和国家安全方面开创新的成功。”

在报告中，两个委员会还例举了美国癌症研究所以及中央情报局如何通过收集海量数据、建立数据仓库、实施以数据挖掘为核心的自动分析技术，获得了出人意料的创新和成功。

委员会一致认为，如何有效地利用数据将贯穿所有科技领域的挑战。最后，两个委员会向奥巴马建议：联邦政府的每一个机构和部门，都需要制定一个“大数据”的战略。

大数据（Big Data）

大数据是指那些大小已经超出了传统意义上的尺度，一般的软件工具难以捕捉、存储、管理和分析的数据。

但是，具体多大的数据才能称为“大”，并没有普遍适用的定义。一般认为，大数据的数量级应该是“太字节”（$2^{40}$）的。麦肯锡全球研究所认为，我们并不需要给“什么是大”定出一个具体的“尺寸”，因为随着技术的进步，这个尺寸本身还在不断地增大。此外，对于各个不同的领域，“大”的定义也是不同的，无需统一。

其实，“大数据”这个名词并不新鲜，早在1980年代，美国就有人提出了“大数据”的概念。20多年来，各个领域的数据量都在迅猛增长，美国的企业界、学术界也不断地对这个现象及其意义进行探讨，“大数据”这个名词变得越来越流行、越来越重要，最后成为了国家和政府层面的发展战略。

之所以要称之为战略，是因为“大数据”之“大”，并不仅仅在于其“容量之大”。当然，由于数据容量的爆炸，数据的收集、保存、维护以及共享等等任务，都成为具有研究意义的现象和挑战。但“大数据”之“大”，更多的意义在于：人类可以“分析和使用”的数据在大量增加，通过这些数据的交换、整合和分析，人类可以发现新的知识，创造新的价值，带来“大知识”、“大科技”、“大利润”和“大发展”。

如前文所述，数据，是记录信息的载体，是知识的来源。数据的激增，意味着人类的记录范围、测量范围和分析范围在不断扩大，知识的边界在不断延伸。

2007年，雅虎的首席科学家沃茨博士在《自然》上发表了一篇文章《21世纪的科学》[22]，他发现，得益于计算机技术和海量数据库的发展，个人在真实世界的活动得到了前所未有的记录，这种记录的粒度[23]很高，频度在不断增加，为社会科学的定量分析提供了极为丰富的数

据。由于能测得更准、计算得更加精确，他认为，社会科学将脱下“准科学”的外衣，在21世纪全面迈进科学的殿堂。例如，新闻的跟帖、网站的下载记录、社交平台的互动记录等等都为政治行为的研究提供了大量的数据，政治学这门古老的学科，将登堂入室，成为地道的“科学”。

麻省理工学院的教授布伦乔尔森（Erik Brynjolfsson）则比喻说，大数据的影响，就像4个世纪之前人类发明的显微镜一样。显微镜把人类对自然界的观察和测量水平推进到了“细胞”的级别，给人类社会带来了历史性的进步和革命。[24]而大数据，将成为我们下一个观察人类自身社会行为的“显微镜”和监测大自然的“仪表盘”。

这个新的显微镜，将再一次扩大人类科学的范围，推动人类知识的增长，引领新的经济繁荣。麦肯锡全球研究所在其2011年的报告中最后概括说：大数据，将成为全世界下一个创新、竞争和生产率提高的前沿。

抢占这个前沿，无异于抢占下一个时代的“石油”和“金矿”。

2012年3月29日，奥巴马政府又进一步推进了其“大数据”战略。奥巴马的高级顾问、总统科学技术顾问委员会（PCAST）的主席霍尔德伦（John Holdren）代表国防部、能源部等6个联邦政府部门宣布，将投入2亿多美元立即启动“大数据发展研究计划”（Big Data Research and Development Initiative），以推动大数据的提取、存储、分析、共享和可视化。霍尔德伦也是哈佛大学肯尼迪政府学院的知名教授，他在讲话中表示：像美国历史上对超级计算和互联网的投资一样，这个大数据发展研究计划将对美国的创新、科研、教育和国防产生深远的影响。

奥巴马则强调联邦政府必须和公司、大学结盟，全民动员（All

Hands on Deck），来应对“大数据”时代的挑战。

人类知识的三大种类与科学的关系

人类所有的知识，可以划分为三个大类：自然科学、社会科学和人文艺术。

自然科学的研究对象是物理世界，讲的是“精确”，丝毫不能含糊，卫星上天、潜艇下海，差之毫厘，就会谬以千里。

社会科学研究的是社会现象，探讨的是人和社会的关系，如经济学、政治学、社会学，它也追求精确，但因为关系到多变善变的人，导致了“测不准”，所以社会科学又被称为“准科学”。

人文艺术则主要包括文学、艺术、哲学，它探讨的是人的信仰、情感和价值，并不强调精确，有时候甚至模糊就是美，所以位于科学的最外围。

在科学的谱系里，社会科学正好介于自然科学和人文艺术之间。

## 注释

01 机器学习（Machine Learning），是人工智能的一个分支，通过在大量数据上运行分析程序，达到让计算机自动学习、积累智能的目的。

02 Designing a Digital Future, Page. xvii, The President's Council of Advisors on Science and Technology, Dec 2010.

03 英语原文为：“Moore's Law has been the name given to everything that changes exponentially in the industry. Another decade is probably straightforward…There is certainly no end to creativity.”—Gordon Moore, February 2003

04 数据来源于美国劳工部对于联邦政府的介绍网页：With about 2.0 million civilian employees, the Federal Government, excluding the Postal Service, is the Nation's largest employer. http://www.bls.gov/oco/cg/cgs041.htm

05 从1996年起，美国联邦政府就认为自己是美国最大的单个信息生产、收集、使用和发布方，见OMB Circular A-130。

06 *Big data: The next frontier for innovation, competition and productivity*, McKinsey Global Institute, May 2011.

07 离散式制造业（Discrete Manufacturing）是指将不同的现成元部件装配加工成较大型系统的行业，例如汽车行业。流程式制造业（Process Manufacturing）是指通过一条生产线将原材料制成可以出售的成品的行业，比如制药。

08 Md.-based intelligence agencies helped track Bin Laden, 05-7- 2011, Scott Calvert, *The Baltimore Sun*.

09 本书第五章第一节将会具体解释这个负担小时的计算方法。数据来源为：Information Collection Budget Of The United States Government，Office of Management and Budget，2010。

10 摩尔1965年发表该定律时，认为这个周期是1年，1975年，他修订为2年；也有人认为，这个周期是18个月。

11 “The capacity of digital data storage worldwide has doubled every nine months for at least a decade, at twice the rate predicted by Moore's Law for the growth of computing power during the same period.”Fayyad, U. and Uthurusamy R., Evolving data mining into solutions for insights, Communications of the ACM, Vol. 45, No. 8, 2002, pp. 28-31.

12 英语原文为：“Smart data structures and dumb code works a lot better than the other way around.”—The Cathedral and the Bazaar, Eric Steven Raymond, 1999

13 英语原文为：“Data is a precious thing and will last longer than the systems themselves.”— Tim Berners-Lee, www2006 conference BCS interview, 2006

14 Uncle Sam's first CIO, *Fortune Magazine*, Interview by Geoff Colvin, July 13, 2011.

15 英语原文为：“What I want to get done is what the people desire to have done, and the question for me is how to find that out exactly.”— Abraham Lincoln

16 Election Polls—Accuracy Record in Presidential Elections, *Gallup*, http://www.gallup.com/poll/9442/election-polls-accuracy-record-presidential-elections.aspx.

17 英语原文为：“The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.”—The Computer for the 21st Century, Mark Weiser, 1991.

18 A National Operational Wave Observation Plan, NOAA and USACE, March 2009.

19 Commercial Fishing Fatalities — California, Oregon, and Washington, 2000-2006, Centers for Disease Control and Prevention, April 2008.

20 原文为：“Every federal agency needs to have a ‘Big Data’ strategy.”—Designing a Digital Future, Page. xvii, The President's Council of Advisors on Science and Technology, Dec 2010.

21 3D Data Management: Controlling Data Volume, Velocity and Variety, Laney, Douglas. Feb 2001.

22 A twenty-first century science, Nature 445, 489; 1 February 2007, Duncan Watts.

23 数据粒度是指数据的细化程度。细化程度越高，粒度就越小；相反，细化程度越低，粒度级就越大。

24 The Age of Big Data, *New York Times*, Steve Lohr, February 11, 2012.

# 第三章　数据治国

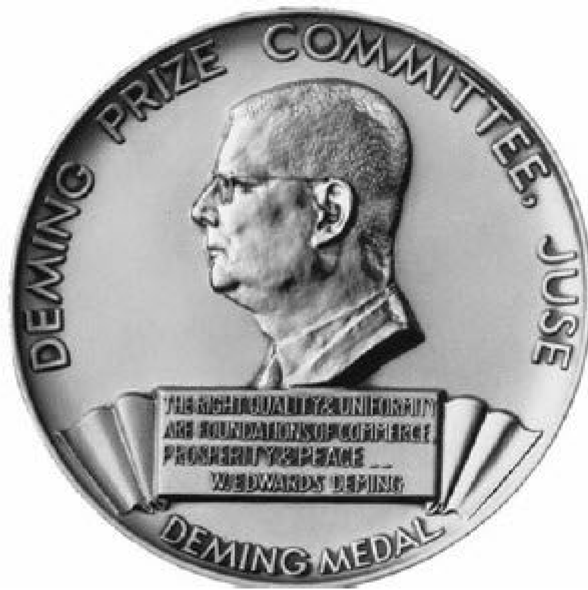
尽管信息时代的技术进步已经彻底改变了商业领域和体育运动领域的决策过程，但联邦政府对这些新技术的应用还仅仅处在一个开始的阶段。要实施“数据驱动的决策方法”（Data-Driven Decision Making），我们不仅要使用新的技术、还要改变目前的决策过程。

然而，一旦打破这些障碍，我们将从中获益：政府将更有效率、更加开放、更加负责，引导政府前进的将是“基于实证的事实”，而不是“意识形态”，也不是利益集团在政府决策过程中施加的影响。

机遇就在眼前。我们现在需要具有远见的政治领袖立刻抓住这个机遇。[01]

——丹尼尔·埃斯蒂，耶鲁大学法学院教授，2007年4月

美国这个国家，虽然年轻，但相信数据、使用数据，却有着深厚的传统。数据被视为科学的度量、知识的来源；没有数据，无论是学术研究，还是政策制定，都寸步难行。




爱德华·戴明（1900-1993）

在1947年之后开始加入日本的经济重建工作，他在统计和质量管理方面的贡献，被认为是日本创新、打造高质量产品的重要助推器，他因此被日本人视为对其经济崛起作出重要贡献的英雄。

1950年，日本科学家和工程师协会开始设立戴明奖；每年的颁奖典礼，国家电视台都现场直播，为年度盛事。图为戴明奖奖牌。

爱德华·戴明（Edwards Deming）是享有世界声誉的美国统计学家、管理学家，也是质量管理理论的奠基人。二战结束后，他曾经旅居日本，帮助日本人开展战后重建，为日本的经济崛起立下了丰功伟绩。他有一句名言说：

“我们信靠上帝。除了上帝，任何人都必须用数据来说话。”

这句话在美国流传极广。

“我们信靠上帝”源于美国国歌，印在货币上不说，还是美国的国家箴言（National Motto）。作为一个宗教大国，美国有90％至92％的人都声称相信上帝。[02]“除了上帝，任何人都必须用数据来说话”，这后半句把数据提到了和上帝相提并论的高度，这反映了美国人对于数据的无比推崇。

上帝，是极大、极高、极虚的神的化身；数据，是至小、至实、至真的逻辑单元。既信奉上帝，又推崇数据，这两者貌似对立，但却在美国大众的价值观中交融渗透，这也反映了美国人对于宗教和科学的复杂心情和矛盾态度。

可以想象，随着大数据时代的不断深入，美国人对于“上帝”的态度，可能会越来越纠结；而对“数据”，将会越来越“迷信”。

这种“迷信”，对联邦政府而言，已经不仅仅停留在“用数据来说话”的层次上了。近年来，随着大数据的迅猛增加，各个政府部门都在尝试“用数据来决策”、“用数据来管理”、“用数据来创新”，在这个过程中，涌现了一大批既务实管用、又令人耳目一新的做法和应用。

## 循“数”管理：平安大道怎样铺

截至2010年，中国拥有7800万机动车辆，大概相当于美国的三分之一。但根据中国警察部门的报告，2010年有7万人死于交通事故，是美国的两倍。真实的数据可能还要高很多。两国的交通安全专家认为：美国的数据非常可靠，统计了每一宗死亡案例，但中国的地方警察部门普遍存在少报、瞒报的问题，只有一小部分交通死亡的数据最终进入了官方的报告。[03]

——《纽约时报》，2011年7月26日

美国是全世界第一个全面普及汽车的大国，一度被称为“车轮上的国家”。

汽车，是19世纪末在欧洲发明的。但一进入20世纪，美国就开始在汽车领域领跑世界。这主要归功于一个美国人：亨利·福特（Henry Ford）。

福特出生于一个农庄家庭，他的父亲寄望于他继承农场的事业，但他却喜欢机械发明，15岁就自己装配了一台内燃机。16岁的那年，福特离开了家乡，去工厂学徒、打工，他最终如愿以偿，成了一名机械工程师。1896年，33岁的福特自己设计制造了第一辆汽车。

后来，他的车开得越来越快，他也在这条路上，走得越来越远。

1903年，他创立了福特汽车公司。在接下来的25年内，其公司为美国生产了1500万辆汽车。

福特对美国社会的杰出贡献，不仅仅在于他是一个发明家、企业家，还在于他具有极强的社会责任感。他立志要让汽车这个当时只有权贵人物才能享用的奢侈品走进普通家庭，率先在汽车制造领域引进了流水线装配的大规模生产方式。他曾经说道："我将为大众生产汽车。但它的价格会很低，只要有一份像样工资的人，都能买得起，在上帝开创的伟大空间里，他和他的家庭能共享快乐的时光。"[04]

在这个理念的指导下，福特让利、低价销售他的产品。通过他的努力，汽车在美国变成了真正的"大众消费品"。福特也被后世称为"美国汽车之父"。

随着汽车的普及，道路交通事故也随之增多，成为一个突出的公共问题。从1960年到1965年，因交通事故而死亡的人数以每年近30％的速度增长；1966年，死亡人数首次突破了5万，成为全美舆论的焦点。当时，专家估计，如果不采取有力的措施来扼制这个增长势头，死亡人数将在1975年达到10万。

但这却没有成为现实。

接下来的几年，死亡人数有升有降。到1972年，交通事故的死亡人数攀升到了54589人，达到了历史的最高点。此后近40年，呈不断下降的趋势。2009年，死亡人数为33808人，创下了自1954年以来的最低水平。不仅绝对死亡数量大幅下降，相对值也降到历史最低水平，每1亿英里驾驶里程的死亡数为1.13人。

这个成绩的取得极为不容易。

要知道，1966年，美国的人口数为1.96亿，2009年，已经3.07亿；1966年，美国仅有1.01亿注册驾驶员、0.94亿机动车辆，2009年，驾驶

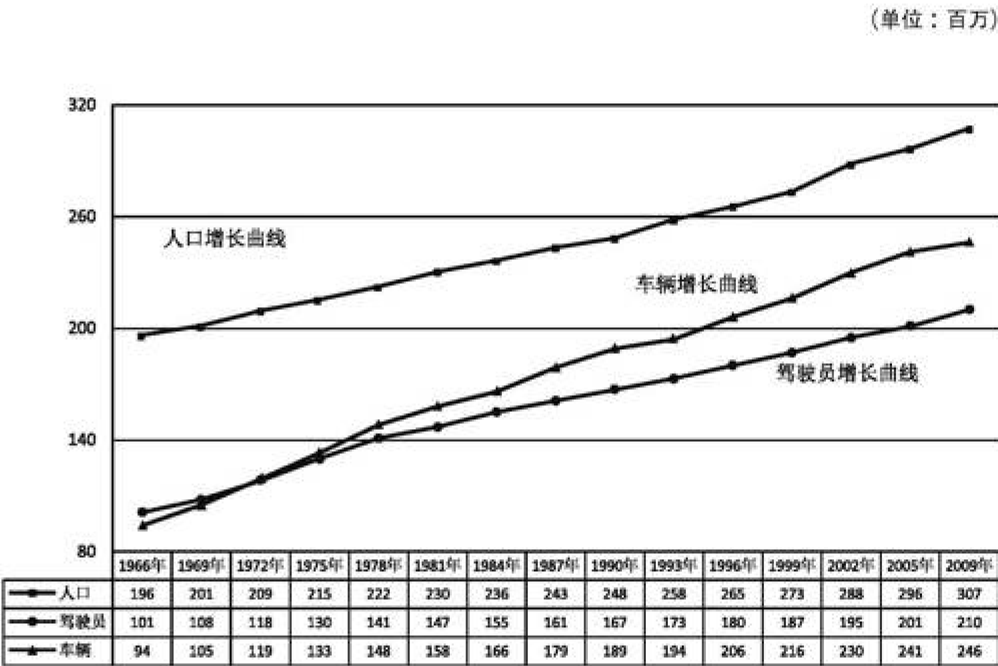
员上升到2.1亿，机动车量增加到2.46亿，人口和车辆的数量、密度都成倍增长，车辆的使用频率也大幅增加，但交通事故的死亡人数却不升反降，而且幅度显著：由5万多人下降到3万多人。

原因当然不是单方面的。

**1966**年和**2009**年的主要指标对比

| 项目 \ 年份 | 1966 年 | 2009 年 | 趋势 |
|---|---|---|---|
| 人口 | 1.96 亿 | 3.07 亿 | ⬆ |
| 车辆 | 0.94 亿 | 2.46 亿 | ⬆ |
| 交通事故死亡人数 | 突破 5 万人 | 约 3.38 万人 | ⬇ |

**1966**年到**2009**年美国人口、机动车辆以及驾驶人员的增长




| | 1966年 | 1969年 | 1972年 | 1975年 | 1978年 | 1981年 | 1984年 | 1987年 | 1990年 | 1993年 | 1996年 | 1999年 | 2002年 | 2005年 | 2009年 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 人口 | 196 | 201 | 209 | 215 | 222 | 230 | 236 | 243 | 248 | 258 | 265 | 273 | 288 | 296 | 307 |
| 驾驶员 | 101 | 108 | 118 | 130 | 141 | 147 | 155 | 161 | 167 | 173 | 180 | 187 | 195 | 201 | 210 |
| 车辆 | 94 | 105 | 119 | 133 | 148 | 158 | 166 | 179 | 189 | 194 | 206 | 216 | 230 | 241 | 246 |

说明：这3条曲线表明，美国机动车辆的增长速度比人口的增长速度、驾驶人员的增长速度都快；1970年代，机动车辆的总数开始超过驾

驶人员的总数。［数据来源：美国高速公路管理局（FHA）］

一方面，可以肯定，因为汽车技术的发展，汽车本身的安全性在不断提高；但另一方面，交通事故的发生，很大程度上是人为原因造成的，死亡人数之所以能够大幅下降，更重要的原因，在于政策的引导和管理。

美国的交通安全管理局（NHTSA）是交通事故的主管部门。回顾这几十年的工作，该局将他们的经验概括为：循“数”管理。

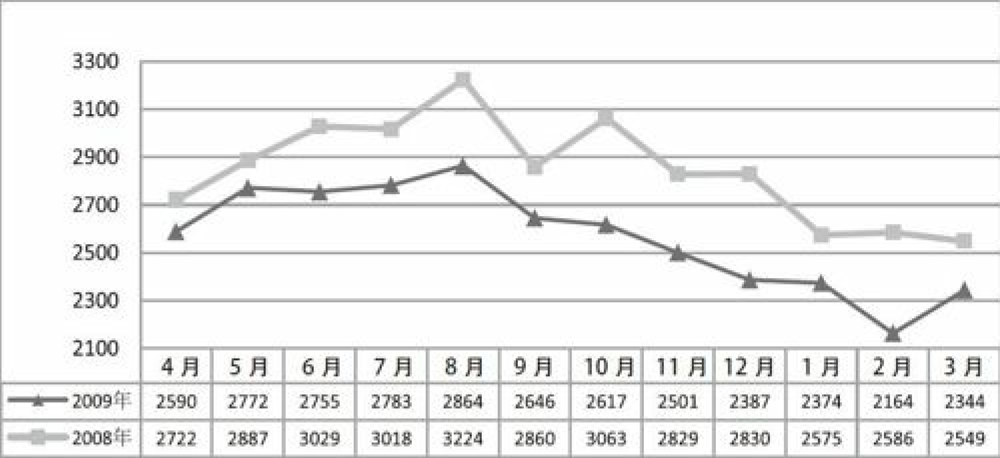
循“数”管理的方法，就起源于1966年。

1966年，由于交通事故死亡人数突破5万人，民间一时舆论沸腾，美国国会迅速对此作出了回应，通过了《高速公路安全法》（Highway Safety Act），要求联邦政府“立即建立一套有效的交通事故记录系统，以分析确定交通事故及伤亡的原因”。[05]

这个法案的直接结果，是交通安全管理局开始在全国范围内收集交通事故的死亡记录，建立了“交通事故死亡分析报告系统”（Fatal Analysis Reporting System）。交通安全管理局也因此成为美国联邦政府最早开始大规模收集数据的部门之一。经过几十年的发展，该系统已经演变为一个在线分析系统，任何人都可以上网查询。

以2009年为例，在该局的网页上，你可以轻松找到以下数据，[06]仅仅做一些简单的分析和对比，你就能发现，美国发生的交通事故确实有“规”可循：

**2008、2009**年致命交通事故按月对比




发现［1］：从两年的数据看，夏季（5、6、7、8月）明显是一年之中交通事故的高发期。

**2009**年致命交通事故按星期和时间对比

| | 星期天 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 | 星期六 | 总计 |
|---|---|---|---|---|---|---|---|---|
| 0:00-3:00 | **1072** | 316 | 270 | 328 | 422 | 525 | **1024** | 3957 |
| 3:00-6:00 | 635 | 245 | 221 | 259 | 276 | 337 | 614 | 2587 |
| 6:00-9:00 | 324 | 451 | 388 | 400 | 428 | 429 | 378 | 2798 |
| 9:00-12:00 | 392 | 420 | 409 | 373 | 395 | 452 | 493 | 2934 |
| 12:00-15:00 | 559 | 572 | 542 | 493 | 524 | 622 | 669 | 3981 |
| 15:00-18:00 | 700 | 687 | 677 | 663 | 659 | 767 | 791 | 4944 |
| 18:00-21:00 | **768** | **627** | **594** | **628** | **680** | **807** | **884** | **4988** |
| 21:00-24:00 | 523 | 448 | 484 | 540 | 592 | 892 | 880 | 4359 |
| 时间不明 | 31 | 31 | 30 | 40 | 28 | 36 | 53 | 249 |
| 总计 | **5004** | 3797 | 3615 | 3724 | 4004 | **4867** | **5786** | 30797 |

发现［2］：每天的18:00-21:00是交通事故的最高发时段，但可以看到，周六这个时段的事故发生量比工作日还多。这说明，不仅仅是因为上下班，即使休息，18:00-21:00这个时段也是人们驾车出行最活跃的时段。

发现［3］：单就一天来看，黑色星期六为“单天”死亡数之冠，其次是星期天、星期五。

发现［4］：如果仅仅从一周内的时段来看，星期天的午夜零点至3点，才是真正的致命时段，其次是星期六的同一时段。

**2009**年致命交通事故按时间段对比

| | 单车事故 | | | 多车相撞 | | | 合计 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 数量 | 酒后驾车 | 酒后驾车比例 | 数量 | 酒后驾车 | 酒后驾车比例 | 数量 | 酒后驾车 | 酒后驾车比例 |
| 0:00-3:00 | **3099** | 2030 | **66%** | 858 | 568 | **66%** | 3957 | 2598 | **66%** |
| 3:00-6:00 | 1935 | 1042 | 54% | 652 | 304 | 47% | 2587 | 1346 | 52% |
| 6:00-9:00 | 1536 | 301 | 20% | 1262 | 128 | 10% | 2798 | 429 | 15% |
| 9:00-12:00 | 1412 | 159 | 11% | 1522 | 94 | 6% | 2934 | 253 | 9% |
| 12:00-15:00 | 1898 | 267 | 14% | 2083 | 185 | 9% | 3981 | 452 | 11% |
| 15:00-18:00 | 2474 | 594 | 24% | 2470 | 363 | 15% | 4944 | 957 | 19% |
| 18:00-21:00 | 3104 | 1044 | 34% | 1884 | 585 | 31% | 4988 | 1629 | 33% |
| 21:00-24:00 | 3056 | 1401 | 46% | 1303 | 623 | 48% | 4359 | 2024 | 46% |
| 时间不明 | 231 | 120 | 52% | 18 | 6 | 32% | 249 | 126 | 50% |
| 总计 | **18745** | 6958 | 37% | **12052** | 2856 | 24% | 30797 | 9814 | 32% |

发现［5］：午夜零点至3点这个时段发生事故的原因，有66％是酒后驾驶；在所有的时段当中，午夜酒后驾驶的比例是最高的。

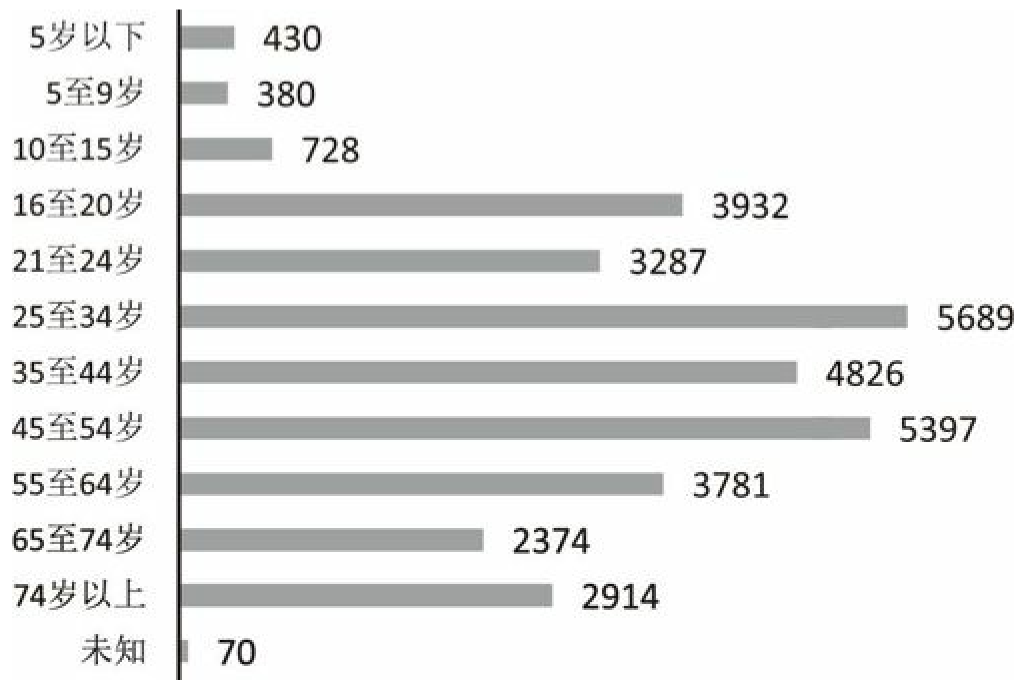
发现［6］：单车驾驶产生的事故远远高出多车相撞的事故。

**2009**年致命交通事故按天气和路况对比

| | 白天正常光线 | 晚上有灯光 | 晚上无灯光 | 黎明或黄昏 | 情况不明 | 总计 |
|---|---|---|---|---|---|---|
| 正常天气 | **13408** | 4811 | 7813 | 1037 | 55 | **27124** |
| 雨天 | **1055** | 486 | 747 | 115 | 7 | 2410 |
| 雪天 | **307** | 43 | 207 | 34 | 2 | 593 |
| 其他 | **113** | 51 | 200 | 33 | 0 | 397 |
| 天气不明 | **65** | 14 | 116 | 3 | 75 | 273 |
| 合计 | **14948** | 5405 | 9083 | 1222 | 139 | 30797 |

发现［7］：恶劣天气和光线条件并不是发生事故的必然原因，绝大部分事故都发生在正常天气及光线较好的白天。

**2009**年因交通事故死亡人员的年龄结构



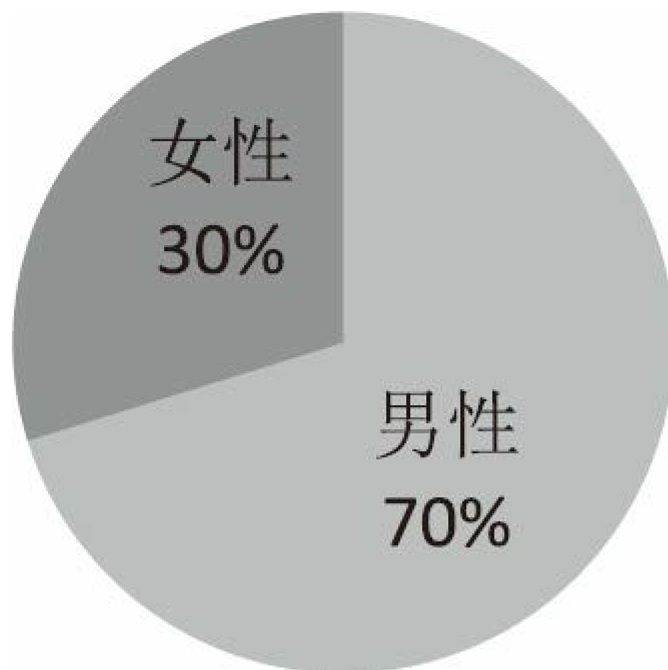

发现［8］：25岁到34岁的人群居交通事故死亡人数之首，其次是45岁至54岁的人口群。

**2009**年致命交通事故死亡人数按人员类别对比

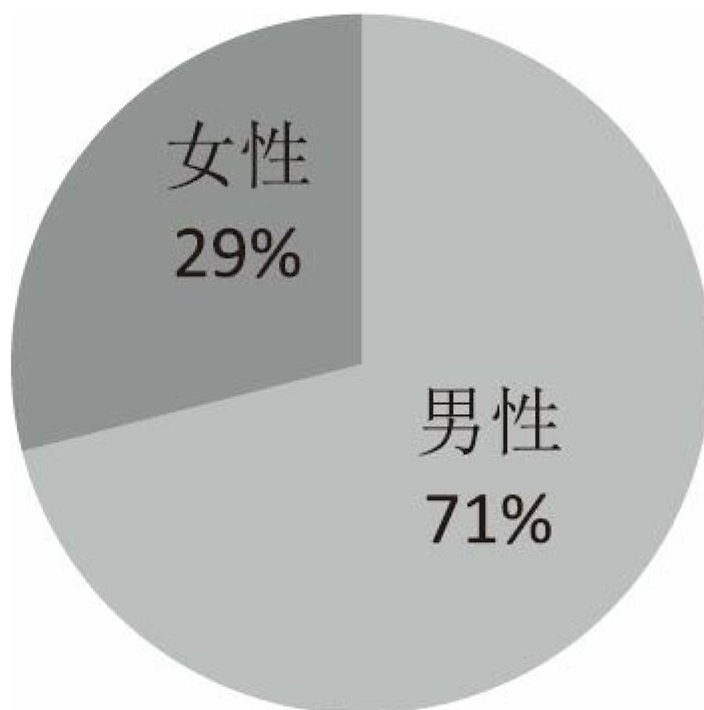
| | 人员类别 | 数量 | 小计 |
|---|---|---|---|
| 汽车 | 汽车司机 | 17640 | 24474 |
| | 乘客 | 6770 | |
| | 不明 | 64 | |
| 摩托车 | 摩托车司机 | 4462 | 4462 |
| 行人 | 行人 | **4092** | 4872 |
| | 骑自行车的人 | 630 | |
| | 其他 | 150 | |

发现［9］：行人总是交通事故的牺牲品，约占死亡人数的14.4％。

**2009**年因交通事故死亡的男女比例




**2008**年因交通事故死亡的男女比例




发现［10］：男女两性因交通事故死亡的比例很稳定，为7:3。

以上的图表和发现，都是仅仅基于一两年数据之上的简单加总、分析和对比。

可以想象，随着数据的累积和增多，可以做的分析和对比也越来越多。一起交通事故的数据可能是无序的，一年的数据、一个地区的数据也看不出太多的章法，但随着跨年度、跨地区的数据越来越多，群体的行为特点就会在数据上呈现出一种“秩序、关联和稳定”，更多的规律将浮出水面。在几十年的循“数”管理中，美国国家交通安全管理局有不少这样的经验和例子。

有一次，该局发现某个州发生车辆右侧碰撞的比例每年都比其他州高，在认真调查之后，发现原因在于该州公路的路缘坡比其他州的都要长，这导致了驾驶员注意力的分散。又比如，1980年始，个别州开始实施驾驶人员必须佩戴安全带的规定，但随后收集到的数据却表明，实施同样规定的州，死亡率的下降幅度却各不相同。这个数据之差引发了联邦政府对全国的执行情况进行检查和对比。后来发现，效果明显的州，警察有权力随时截停车辆、检查司乘人员是否佩戴了安全带，而效果不明显的州，仅仅规定警察只能在以其他理由截停车辆时顺便检查车上的人员是否佩戴了安全带。这个发现，后来引发了多个州执法方式的改变。这也说明，相同的政策，由于执行方式不同，效果可能大不相同。

近几十年以来，交通安全管理局每年都组织各州的工作人员参加数据收集和分析的培训。根据数据分析的结果，该局调整、制定新的政策，新政策实施以后，再收集新的数据，进入新一轮的效果评估，如此周而复始、循“数”渐进，从而确定最有效的措施、最好的做法，再在全国推广。

除了数据收集和数据分析之外，该局的经验表明，循“数”管理还有

一个重要的环节：数据发布。

交通安全管理局通过互联网公开发布其收集的数据。

如果认为网上的数据发布仅仅是为了社会监督，那就大错特错了。监督仅仅只是其中一个目的。交通安全管理局在介绍他们循“数”管理的经验时，特别强调了通过网络发布数据的重要意义。

在网上发布数据，将会吸引一大批对这个问题感兴趣的各界人士参与到政策制定的过程中来，足够多的眼睛，将会使所有的问题都无所遁形，更多的问题将被发现，更多的细节将被讨论，更好的方案将会被激发。还有，通过发布这些数据，一项具体政策的执行效果能够被量化，各个州之间很方便进行“事实对比”，新的政策如果有效，很容易在全国范围内被接受、复制并推广，形成一种良性的“政策竞争”氛围。

交通安全管理局的循“数”管理三部曲



# 数据“验”平权：民权史上的碑石

（但对黑人来说）仅仅自由是不够的。你不能只对他们说：“现在你自由了，想去哪里都可以，也可以做你想做的事，选择你喜欢的领导人。”仅仅这样做不可能抹去几百年的创伤。一个多年来都戴着枷锁蹒跚的人，你不能一解除他的桎梏，就把他带到起跑线上，对他说：“你自由了，现在可以和其他人竞争了。”你还认为这是真正的公平。……我们追求的不仅仅是自由，还有机会……就此而言，机会平等才是根本。[07]

——林登·约翰逊，第36任美国总统

1965年6月4日在霍华德大学毕业礼上的演讲

在美国，和政府做生意的公司有一个专门的名称，叫政府合同商（Government Contractor）。政府合同商通过承接政府的项目，赚国家的钱。国家的钱，归根结底就是纳税人的钱。为了保证纳税人的钱不仅花得有效率，而且还能够促进社会公平，联邦政府对这个过程制订了诸多的规定。例如，对于50人以上的公司，只要获得一笔大于5万美元的政府合同，联邦政府对其就执行有别于一般企业的要求。诸多要求中的重要一项，就是“平权措施计划”（Affirmative Action Plan）。

平权措施计划，是美国民权运动史上的里程碑。

该计划规定：公司不得因为雇员的年龄、种族、肤色、原籍、性别、婚姻状态、宗教信仰，以及是否残疾、是否退伍军人等事项对雇员产生歧视。其初衷在于保证所有人，特别是少数民族，都享有公平就业

的机会，在每一个公司都营造一个和总体人口特征一致的工作集体。

换一个角度来思考，“平权措施”其实并不“平”，因为其本质是通过人为的干预给予少数民族“优先”权，制造机会平等。比如说，一个公司要增加人手，即使前来应聘的白人要比黑人优秀，但因为法定的黑人比例，可能仍然要雇用黑人。

但很多人认为，只有通过这种平权措施，才能在现阶段有效地帮助少数族裔，特别是缩小黑人与白人之间的差距。

“平权措施计划”起源于1961年的肯尼迪时代，那时候的平权，主要集中在种族和肤色的问题上。1963年，肯尼迪遇刺，约翰逊继位，这位和《信息自由法》过不去的总统，极力推行种族平等，将“平权措施”推向了顶峰。1967年，受女性解放运动的影响，他又将平权措施扩大到“性别”领域。近10年来，在美国，“性别平等”甚至有超过“种族平等”的趋势，成为平权措施计划中最受争议的话题。

2001年，沃尔玛公司150万女雇员认为她们在公司的薪酬和晋升中受到了普遍的歧视，从而提起了性别歧视的集体诉讼。该案原告人数之多，创下历史之最。官司历时10年，一直从联邦地方法院、巡回法院打到最高法院，“百万娘子军”曾两度胜诉，但2011年6月，最高法院判决，原告缺乏特定的、直接的证据证明沃尔玛存在普遍的性别歧视政策，不予支持。

“娘子军”虽然最终败诉了，但女性在职场的平等问题已经成为一把随时都可能落下来的“达摩克利斯之剑”，作为联邦政府的合同承包商，各大公司无不警醒、小心翼翼。

那么，联邦政府的劳工部是如何监督落实这些平权政策的呢？

每一年，每一个政府的合同商都要上交劳工部两份报告，一是数据统计报告，二是文字叙述资料。数据统计报告中包括了全部员工的性别、种族以及收入变化等数据项，劳工部的工作人员对每个公司提交的数据进行分析，并和行业的总体情况进行对比，一旦发现潜在的歧视问题，例如女性职员的数量过少或薪酬过低，就会对相应的公司发出警告，限期改正，拒绝改正的公司将失去再次获得合同的资格。文字叙述资料则包括公司的员工手册、宣传相关劳工法的海报、培训制度、招聘制度以及相关职位招聘过程的记录。

2010年，全美共有30多万个政府合同商，由于上述这项规定，每年所产生的数据量，也不容小觑。基于原始数据的分析，当然也更加准确、公正、可靠。

政府合同商向联邦政府提交的这些数据和信息，当然也属于联邦政府收集的业务信息。那么，除了劳工部把它们用于分析监督“平权措施计划”的实施情况之外，是否可以公开呢？换句话说，如果有第三方按照《信息自由法》对联邦政府提出要求，要查阅某政府合同商为“平权措施计划”提交的信息和资料时，联邦政府是否可以公开这些资料呢？

这个问题曾经在美国历史上“纠结”过，引起过很多宗法律纠纷。其中最著名的是“克莱斯勒诉布朗案”。[08]

1970年代，克莱斯勒公司（Chrysler Corp.）获得了国防部的一项合同。作为政府合同商，该公司按规定向国防部提交了“平权措施计划”规定的资料。1975年5月，一个工会组织向国防部提出信息查询的申请，要求获得克莱斯勒公司提交的数据和文件。国防部认为该要求符合《信息自由法》的规定，在知会克莱斯勒公司后，向该工会公开了这些数据和文件。

但克莱斯勒公司对此表示强烈的反对，他们认为这些信息属于它的商业秘密，国防部无权向第三方公开。克莱斯勒随即将当时的国防部部长布朗告上了联邦地方法院，经过3年的折腾，1978年，官司最后上诉到最高法院。

1979年4月，最高法院宣布克莱斯勒败诉。最高法院认为，《信息自由法》的根本目的，是鼓励政府机构的信息公开，商务机密虽然享有豁免权，但根据现有的法律，这种豁免权是许可性的，而非强制性的，既然是非强制性的，即使该信息属于商业机密，政府部门有权不公开，也有权公开，即公开并不违法。

克莱斯勒诉布朗案是“平权措施计划”及《信息自由法》的标志性案件。此后，《信息自由法》也开始正式成为商业公司竞争的工具。不少公司，都通过这个渠道，向联邦政府索要文件，以了解竞争对手的运营情况。据统计，2010年，有三分之一的信息公开的要求都来自商业公司。这当然又引起了美国社会的又一次讨论，《信息自由法》的初衷，是保护公民的知情权，如今似乎已经“沦”为商业竞争的工具，这又是否合适呢？

# 数据“打”假：最大的争议就是福利滥用

我们才刚刚开始全面利用数据挖掘的技术。我们相信，使用正确的技术，加上专业的人员，并保证数据安全，通过这三者的结合，数据挖掘将成为我们提高财务审计效率和效果的重要工具。[09]

——美国国会政府问责办公室，关于数据挖掘在国会的听证，2003年3月

欧洲大部分的发达国家都实现了全民医疗保险制度。

但对这种全民福利的做法，长期以来，美国社会并不认同。

不认同的原因主要有两点：一是对政府的不信任；二是认为如果政府对公民大包大揽负全责，公民就会对自己不负责。

在这种态度的主导下，美国的福利制度，一直以来都是以具体的项目为依托、以特定的群体为目标来开展的。其中，两个最大的医疗福利计划是1965年通过的，也是第36任总统约翰逊促成的：

一是专门针对残障人士和65岁以上老人的医疗保险计划（Medicare）；

二是以贫困人口为对象的医疗补助计划（Medicaid）。

也就是说，老、弱、穷看病，政府买单。第一个医疗保险计划是通过保险来支付，由联邦医疗保险和补助中心（CMS）监管实施；第二个医疗补助计划是政府直接支付，由联邦政府CMS中心和各个州政府共同

实施，分成支付。

根据CMS中心的统计，2009年，医疗保险计划支出共5023亿美元，覆盖了4700万的美国人口；医疗补助计划共支出3739亿美元，覆盖了5680万美国人口。[10]由于两个计划都是针对社会的弱势人群，它们之间也互有交叉，例如，对于老年的贫困人口群，就可能同时适用两个计划。2009年，有850万人同时受惠于两个项目。

福利政策在全世界最大的争议，就是福利滥用、最后养了懒人，导致社会发展缺乏动力。美国虽然慎之又慎，但这两块免费大蛋糕，也不例外。2008年，联邦调查局在《财务犯罪年度公开报告》中估计，联邦政府每年的医疗开支当中，大概有3％到10％涉嫌造假和欺诈。[11]面对利益，人类的创造力总是无穷的：虚假账单、重复申报、无中生有，小病大治、慢治、长治，隐瞒收入和存款、装穷吃低保，林林总总，不一而足。

两个项目的受惠人群，加起来有近1亿人，平均每人每月上一次医院，一年就有12亿张账单，按照联邦调查局的估计，大概会有3600万到1.2亿张账单存在问题。如果要依靠人力去张张审查，其工作量是难以想象的。

目前，CMS中心最主要的措施，就是通过数据来打假。

2001年，加州州政府率先推出了一个数据挖掘的项目“保险补助双向核对”（Medical-Medicaid Data Match），将医疗保险和医疗补助两个项目的数据整合起来，利用两个计划中的人员、时间、价格、地点等数据信息对每一宗申报进行互相核实，通过计算机算法自动确定相互矛盾、有异于常态的支付记录，一旦发现造假或者不实申报的可疑账单，

则转入人工追讨的环节。

这大大缩小了人工审查的范围，提高了打假的效率。

由于效果显著，2004年，这个项目在个别州的实施范围进一步扩大，由事后追讨推进到了事前防范。这意味着，CMS中心一收到申请报销的账单，就启用数据挖掘系统对该申请的数据记录进行“风险评分”：得分低于预定风险额度的申报则转入自动支付的环节；得分超过该风险额度的申请将退回申请人，或提交专门的小组进行人工复审。

支付后的账单，最后还要经过另一个数据挖掘算法的审核，发现可疑的，再转入人工追讨的环节。

以上措施的效果非常显著，很快引起了美国国会的重视。大部分国会议员都相信，数据挖掘的做法将为国家节省开支。2005年，国会通过了《赤字削减法案》（Deficit Reduction Act of 2005），该法案的根本目的是“精简机构、裁减开支”。在对大多数联邦机构削减经费的情况下，国会却给CMS中心下拨了专款，用于成立扩大专门的数据挖掘审查队伍。卫生部甚至主张，这种数据打假的项目应该纳入国家医疗信息系统基础设施建设的范围，换句话说，以后的医疗系统，在设计阶段，就应该包括数据挖掘的打假功能。



说明：2010年10月6日，奥巴马和Technology CEO Council的成员座谈，商讨如何利用高端的数据分析技术来削减开支。（图片来源：www.techceocouncil.org）

数据挖掘还常常能给工作人员带来一些意外的惊喜。例如，Neulasta是一种治疗癌症的注射剂，2006年，CMS中心的算法发现，不同医院提交的Neulasta账单不一致，有的账单使用毫克作为单位，有的账单使用瓶作为单位，而一瓶为6毫克。经过调查证实，使用瓶的账单属于“有意或无意”的人为错误，但这个错误，把申报的开支扩大了6倍。仅该项发现，就为联邦政府挽回了近50万美元的损失。

2007年3月，CMS中心的首席财务官希尔（Timothy B. Hill）出席国会的听证会，他例举了数据挖掘项目实施一年多之后的效果：通过数据挖掘预设的风险评分门槛，确定了2500万美元的虚假申报，事后的数据挖掘复审环节又追讨了1500万美元的超额申报；其中，有50多宗欺诈案

浮出水面，最后移交司法部门处理。[12]

除了CMS中心，联邦政府的社会福利项目都陆续开始采用数据打假的做法。2010年10月，为了削减赤字，奥巴马在白宫和他的技术顾问委员会召开会议，IBM、DELL、Intel等大公司的CEO向他建议说，联邦政府如果在社会福利的项目上加大数据挖掘、分析技术的应用，预计10年内可以再为国家节省2000亿美元。

近十几年来，除了福利打假，数据挖掘的技术在联邦政府的其他部门也获得了广泛的应用。“9·11”以后，国家安全局更是依赖该项技术跟踪确定恐怖分子，本书后续章节对这项技术及其应用还将有专门的介绍。

## CompStat：街头警察的创新传奇

每一个人的心里，都有理想主义的火花，它可以被燃成火焰，进而迸发出非凡的能量和结果。[13]

——路易斯·布兰代斯（1856-1941），美国最高法院大法官

将数据分析大规模地引入到治安管理工作中的做法，起源于纽约。

纽约，是全世界的金融和商业中心，在美国的经济和政治生活当中占有举足轻重的位置。

纽约市也是美国人口数量最多、密度最大、多元化程度最高的城市。它拥有810多万人口，其中超过三分之一是非美国本土出生的外籍人员，这些人来自全世界不同的国家和族裔，使用100多种语言。

因为人口众多，不免鱼龙混杂，纽约也曾经是一个著名的犯罪之都。从上个世纪70年代起，黑帮横行、毒品泛滥，该市的治安情况不断恶化。1990年，纽约市共发生了凶杀案2245宗、车辆盗窃案147123宗，平均每天有6个人死于恶性犯罪，每小时有16台车辆不翼而飞。

1994年，纽约市的警察部门启用了一个新的治安信息管理系统。这是一个以地图为基础的统计分析系统，随着它的出现，纽约城的治安开始逐年好转。这个叫做“CompStat”的系统也开始名扬全国，成了20世纪美国警务管理工作当中最为浓墨重彩的一笔。

CompStat，是COMPuter STATistics（计算机统计）的缩写，现在已经演变成为一个专有名词，特指一种警务管理模式。

这是一个传奇故事，一个源于地铁，发生在一位巡警、一位局长和一位市长之间的传奇故事。

1970年，杰克·梅普尔（Jack Maple）刚刚高中毕业，他加入纽约市交通警察局成为一名地铁线上的警察。当时，地铁线上的抢劫案非常频繁，地铁警察，被认为是纽约最危险的工作之一。

为了追踪抢劫案，梅普尔在时代广场做过便衣，在中央车站指挥过拦截和抓捕。在十几年街头警察的经历当中，他慢慢“悟道”：案件发生在哪里警察就出现在哪里，是让罪犯牵着鼻子跑；要控制局面，抓到老鼠，警察一方必须掌握主动，做一只有“预测能力”的猫。

于是，这位高中毕业生开始研究地铁抢劫案的发生规律。

梅普尔在办公室的墙上挂上了几百幅地图，用不同颜色的大头针来跟踪地铁抢劫案发生的时间和地点，分析其中的原因和规律。无数个夜晚，他点着香烟，站在巨大的地图面前，时而举头凝视，时而低头徘徊，揣度琢磨第二天可能发生抢劫的时间和地点。在一阵苦思冥想之后，最后用大头针按下的那个小点，就代表了他第二天的伏击地点。

梅普尔后来晋升为警督（相当于派出所所长），他就采用这种方法来部署和调配他所辖区的警力。他的办公室挂满了地图，被同事戏称为“地图墙”，他却称之为“预测未来的图表”（Charts of the Future）。

1990年，“预测未来的图表”引起了新任局长布雷特（William Bratton）的注意。布雷特是位退伍军人，他雷厉风行、慧眼识才，在认真研究了“地图墙”之后，他认为梅普尔的方法很“靠谱”。于是开始在全局推广梅普尔的图表管理方法。

第二年，纽约市的地铁抢劫案下降了27％。

但纽约的整体社会治安并没有好转，除了地铁抢劫案，其他的案件都还居高不下。这更令布雷特相信，“预测未来的图表”确实行之有效。

1993年，治安持续恶化。这个问题，甚至成了纽约市市长竞选当中最热门的话题。

这个话题也成为共和党候选人鲁迪·朱利安尼（Rudy Giuliani）的政治机遇。

朱利安尼出生于纽约市的一个普通劳工家庭。他1968年法学院毕业之后，长期在司法部门工作。1983年，他晋升为纽约市的联邦检察官。在他担任这个职务的6年期间，朱利安尼先后将4000多名嫌疑犯送进了监狱，其中有不少都是轰动全国的黑道大哥、帮会首领。朱利安尼作为公诉人，多次受到黑社会的威胁，但他刚正不阿，始终秉法办案，在新闻界获得了“铁面”的美誉。

作为共和党的市长候选人，朱利安尼的主打牌就是“治安”。他在竞选演讲中介绍说，他的朋友在餐馆等公共场合碰到他，都不敢和他打招呼，因为怕被黑社会盯上、莫名其妙受到报复。他在街上，也经常碰到一些刑满释放的人员，那些人对他做出鄙夷的表情，并当面嘲笑他说：“你没人缘，选不上！”

朱利安尼却呼吁说，如果纽约人民真正想改善治安，就不能让他落选！因为他是改善治安最好的人选——别无其他！

他的演讲，情理并茂，他的铁面形象和诚恳的态度，最终打动了纽约人，得以高票当选。

一上任，朱利安尼就立即任命布雷特为纽约市警察局局长。

而布雷特到任的第二天，就任命梅普尔为第一副局长，并要求梅普尔立即组织开发一套电子版的“预测未来的图表”。

“CompStat”于是诞生了。

1994年，互联网还没有普及。CompStat的工作人员每天通过电话和传真向全纽约76个警区收集数据，再将数据统一录入到CompStat，进行加总和分析。

每周二、周四的早晨7点，布雷特就召集全部警区的指挥官开会。最新发生的案件以圆点的形式出现在各个辖区的地图上，不同颜色代表着不同类型的犯罪，特定位置的成串圆点则表明那里发生了一系列的案件。各个指挥官在这些“绩效指示灯”前面依次陈述自己辖区的情况、对策以及警力的调配，一个回合下来，不少人满头大汗。



1980年代，杰克·梅普尔还是地铁站的一名便衣。

他去世当天，纽约城几乎所有的报纸都刊登了他的传奇故事，评价他是“街头警察的伟大发明家”、“真正的纽约英雄”、“熠熠生辉的普通人”。（图片来源：Richmond Hill Historical Society）

为了保证CompStat的落实和推行，布雷特一共撤换了近三分之二执

行不力的指挥官，可谓“铁腕”。

次年，纽约的犯罪率应声而降。凶杀案从1994年的1561宗下降到1177宗，车辆盗窃案由95420宗下降到72679宗。此后，这些重要的指标不断下降。2009年，凶杀案下降到466宗，创下了50年之最低。这个指标，已经使纽约跻身全美最安全的大城市行列。

当然，布雷特在纽约推行的警务管理模式，其做法并不仅仅局限于“地图和数据”。例如，他非常推崇“破窗理论”（Broken Windows Theory）。该理论认为，一个城市，如果对小的违法行为纵容姑息，不良现象就会被放任、模仿，逐渐扩大、蔓延为成片的犯罪行为。所以，即使对一个窗户玻璃被砸破的投诉，接警中心也要认真记录，并纳入地图上的数据分析。也就是说，地图上的圆点不分案值大小、案情轻重，一律同等对待。

又例如，露宿公园、街头的青年，常常是酗酒、卖淫和吸毒活动的参与者，布雷特要求，不能仅仅是打击、驱散，必须刨根问源：他们从哪里来？面对什么困难？需要什么帮助？只有真正解决了这些问题，那些地图上代表犯罪的“圆点”才能最终被消除，而不是从一个区域转移到另一个区域、在地图上“此”起“彼”伏。

2002年，布雷特又担任了洛杉矶的警察局局长。洛杉矶是美国仅次于纽约的第二大城市。在布雷特任职的6年里，洛杉矶的犯罪率每年都在下降。

朱利安尼因为政绩突出，连任两届市长，并登上了《时代》杂志的封面，成了全国知名的人物。2008年，他宣布参选总统，民意调查表明，他是共和党阵营内知名度最高的候选人。但后来由于离婚等个人原

因，他最终退选。

梅普尔后来辞去了公职，创办了一家咨询公司，帮助全国各地的大小警局实施CompStat系统。2001年，他被确诊为患有结肠癌。他留下遗言说，希望自己的灵车能在黄昏的时候，穿越时代广场和中央车站。他说他知道这正是下班的人潮高峰，交通繁忙，但他希望再去那里一次。多年来，他在那里巡逻，确保交通顺畅和治安不乱，这一次，他开玩笑说，要让大家最后也等他一下。

朱利安尼出席了梅普尔的葬礼，他盛赞梅普尔是美国街头警察中伟大的发明家，为纽约城成为美国最安全的大城市做出了历史性的贡献。

纽约的巨大成功，很快引起了其他地方政府和联邦政府司法部的注意。90年代起，全美各地有近三分之一的治安管理部门都陆续复制引进了CompStat的管理模式。1993年，克林顿任命珍妮特·雷诺（Janet Reno）担任司法部长。雷诺是美国历史上第一位女司法部长，在她的任期内，她大力推行“数据驱动”的管理方法（Data-Driven Management），并不断强调：

“数据和信息是执法工作当中制定战略和决策的基础。”[14]

1996年，CompStat获得了哈佛大学的美国政府创新奖（Innovations in American Government Award）。此后，CompStat甚至进入了美国总统和副总统的视野。

1998年10月1日，副总统戈尔宣布，司法部要会同全国政府改革合作委员会[15]成立专门的工作组，在全国的警务部门研究推广“地图映射和数据驱动的治安管理”（Crime Mapping and Data-Driven Management）。克林顿总统甚至在1999年的国情咨文中，也提到这种

新的方法和系统。

随着时间的推移，越来越多的事实和发现证明，这种方法不仅对治安管理行之有效，还可以推广到其他领域，很多时候，还可以带来出人意料的发现。

2006年，通过把20多年的犯罪数据和交通事故的数据整合到一起，并映射到同一张地图上之后，警务研究人员惊奇地发现，交通事故的高发地带，也正是犯罪活动的高发地带，甚至交通事故的高发时间段，也是犯罪活动的高发时间段。

维护交通安全、打击犯罪活动，这两个职能本来分属于不同的联邦部门。基于这个新的发现，美国国家高速公路交通安全管理局（NHTSA）、国家司法援助局（BJA）和国家司法研究所（NIJ）联合成立了一个“数据驱动的新方法：犯罪和交通安全”的工作组，在马里兰、堪萨斯等州的城市开展联合治理的试点。[16]

试点的重点内容是针对犯罪活动和交通事故，为基层警队建立一套完整、严谨的数据整合、分析的系统。由于情况的波动，1年的数据往往不可靠，必须累积3年的数据作为分析的基础；10万人口以下的城市必须使用4到5年的数据。此外，犯罪活动和交通事故极少发生在完全一模一样的地点；除了收集数据，还要在地图上运用簇群关联的数据显示技术，才能划分确定两种活动频发的共同区域。

确定了共同的“黑点”，就可以将交通警察和治安警察的资源整合到一起，有的放矢，在特定的时间、特定的地点开展联合治理，这不仅能提高警力使用的效率，还可以改善执勤巡逻的效果。

联邦政府工作组的试点取得了非常显著的效果，试点区域的抢劫、

盗窃等犯罪活动明显下降，同时，违规驾驶的罚单明显增多。2008年，为了在更多的地方政府推广这种模式，该工作组总结了试点经验，制订了项目实施的基本原则，其中最重要的一条就是：数据收集和数据分析必须成为基层警务部门的一种文化，这种文化代表着基层警务部门一种管理哲学的改变，这种改变，也是该项目实施过程最大的挑战。

这种基于数据的警务管理模式，也引起了学术界的高度关注，有学者将其概括为“数据驱动的警务管理”或“数据驱动的司法管理”（Data-Driven Policing or Justice）。

## 注释

01 Governing by Numbers: The Promise of Data-Driven Policymaking in the Information Age, Daniel C. Esty, Reece Rushing, April, 2007.

02 这个数据综合了皮尤研究中心（Pew Research）和盖洛普咨询公司（Gallup）2010年问卷调查的结果。

03 英语原文为：“China had 78 million vehicles on its roads at the end of last year, roughly one-third as many as the United States. But China has nearly 70,000 police-confirmed traffic deaths a year, twice the figure for the United States. The actual discrepancy may be even greater. Chinese and Western traffic safety experts say that while the United States figures are extremely reliable and take into account virtually every death, only a small fraction of all traffic deaths in China show up in official figures because of widespread underreporting by the local police.”— Road Safety Problems Pose Dire Threat in China, *New York Times*, July 26, 2011

04 英语原文为：“I will build a car for the great multitude. … But it will be so low in price that no man making a good salary will be unable to own one — and enjoy with his family the blessing of hours of pleasure in God's great open spaces.”—*My Life and Work* (1922), Chapter IV, Henry Ford

05 之所以称为“记录系统”，是受限于当年信息技术的发展水平。其英语原文为：“Effective record system of accidents … to determine the probable causes of accidents, injuries, and deaths.”—P.L. 89-564, section 402 (a). (U.S. Code Title 23)

06 该网页地址为：http://www-fars.nhtsa.dot.gov/Main/index.aspx

07 “But freedom is not enough. You do not wipe away the scars of centuries by saying: Now you are free to go where you want, and do as you desire, and choose the leaders you please. You do not take a person who, for years, has been hobbled by chains and liberate him, bring him up to the starting line of a race and then say, you are free to compete with all the others, and still justly believe that you have been completely fair… We seek not just freedom but opportunity…To this end equal opportunity is essential.”—Commencement Address at Howard University, Lyndon B. Johnson, June 4, 1965

08 Chrysler Corp. v. Brown, Secretary of Defense, 441 U. S. 281 (1979).

09 Testimony on Data Mining Results and Challenges for Government Program Audits and

Investigations, United States General Accountability Office, March 2003.

10 http://www.cms.gov/nationalhealthexpenddata/downloads/highlights.pdf

11 FBI's Annual Financial Crimes Report for 2007, Federal Bureau of Investigation.

12 Medicare Program Integrity, Timothy B. Hill, Testimony Before Subcommittee on Health and Oversight and Investigations, Committtee on Ways and Means, U.S. House of Representatives.

13 英语原文为："There is a spark of idealism within every human being that can be fanned into flame, and will bring forth extraordinary results."—Louis Brandeis

14 *Report of the Task Force on Crime Mapping and Data-Driven Management*, U.S. Department of Justice, Attorney General Janet Reno, July 1999.

15 英文名称为：National Partnership for Reinventing Government，该委员会是克林顿执政时期为在联邦政府推行行政改革而成立的临时组织。

16 Data-Driven Approaches to Crime and Traffic Safety, DDACTS, Data-Driven Approaches to Crime and Traffic Safety Operational Guidelines, National Highway Traffic Safety Administration, August 2009.

# 第四章　商务智能的前世今生

仅仅几年以前，这种被称为“商务智能”的技术，还是大公司的专利。但随着计算机处理器、存储器的价格不断下降和软件质量的不断上升，这种技术成了商业界的主流。大大小小的公司，都收集了前所未有的大量数据。过去，这些数据存储在不同的系统当中，如财务系统、人力资源系统和客户管理系统，老死不相往来。现在，这些系统彼此相连，通过“数据挖掘”的技术，可以获得一幅关于企业运营的完整图景，这被称为：一致的真相（A single version of the truth）。商务智能提高了商业运营的效率，帮助了企业总结发展过程中的模式，并改善了企业预测未来的能力。

信息技术产业把商务智能视为对20世纪上半叶企业会计服务、下半叶计算机服务的一个自然承接，正在争相涌入这个领域。爱森哲、普华永道、IBM、SAP都在这个领域投入巨资。技术平台的提供商甲骨文、Informatica、TIBCO、SAS、EMC也从中赢利。IBM更是相信：随着传感器在城市交通、医疗健康中的应用，商务智能将成为其业务增长的顶梁柱。[01]

——《经济学人》，2010年2月25日特别报道

联邦政府这个数据帝国，虽然拥有的数据比任何公司、企业都多，但和私营领域相比，在信息技术的应用上，还是明显落后一步、慢了几拍。

2009年3月，奥巴马就任后的第二个月，就在联邦政府之内设立了

一个全新的职位：首席信息官（Chief Information Officer），并任命来自印度的移民昆德拉（Vivek Kundra）为第一任联邦政府首席信息官。昆德拉在公共和私营两个领域都有广泛的经历，他走马上任之后，曾发表过第一感受：联邦政府信息技术的装备和应用，和一流的商业公司相比，就好像手摇电话摆在了线条圆润、光彩照人的苹果手机旁边，不可同日而语。

当然，这并不奇怪。现代政治学的基本常识告诉我们：由于无法引入有效的竞争机制，政府注定难逃低效的命运。美国联邦政府也不例外。

收集数据、分析数据、发布数据，这一系列和数据有关的信息技术，在商业界其实有个更时尚、更响亮的名字：商务智能。

在商务智能的技术大潮当中，美国联邦政府的做法，只是几朵小小的浪花，公司、大学才是这个领域真正的弄潮儿和领航人。

这股技术浪潮，也在美国起源。

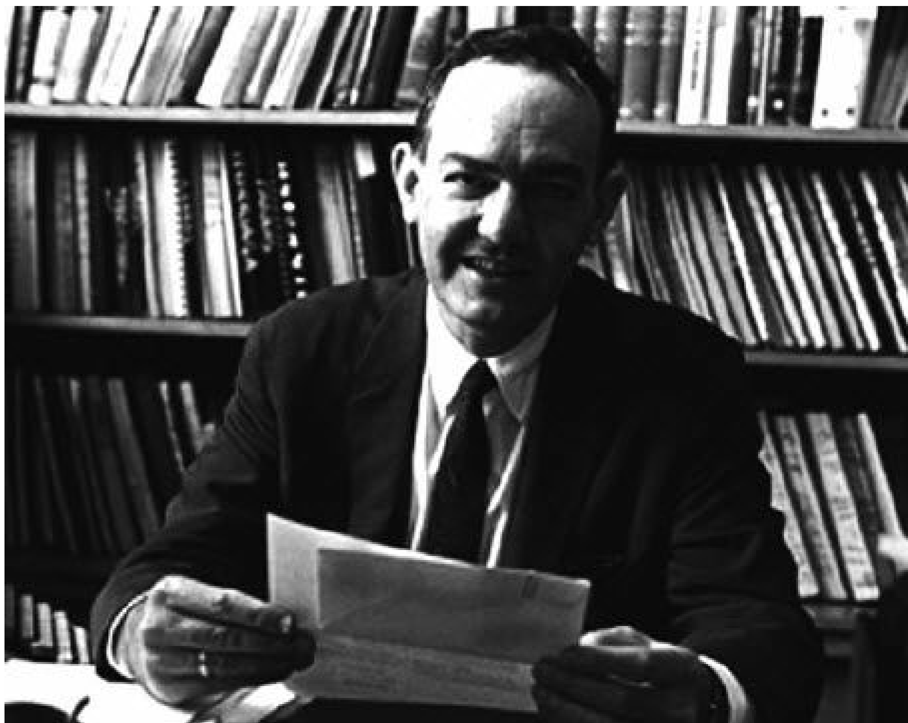
# 起源：从数据到知识的挑战和跨越

信息消费了什么是很明显的：它消费的是信息接受者的注意力。信息越丰富，就会导致注意力越匮乏……信息并不匮乏，匮乏的是我们处理信息的能力。我们有限的注意力是组织活动的主要瓶颈。[02]

——赫伯特·西蒙

美国经济学家、政治学家、人工智能的创始人之一，1973年

1946年，人类历史上第一台电子计算机在美国费城问世。[03]来自匈牙利的移民冯·诺伊曼是这台计算机的主要设计者，他被后世称为“计算机之父”。



赫伯特·西蒙（1916-2001）

20世纪全世界最具影响力的科学家之一，他横跨多个学科和领域，曾获得1975年的图灵奖、1978年的诺贝尔经济学奖、1993年的美国心理协会终身成就奖。（图片来源：卡内基梅隆大学图书馆）

仅一年之后，卡内基梅隆大学的赫伯特·西蒙（Herbert Simon）教授出版了《行政组织的决策过程》一书。在这本被后世视为经典的著作里，他指出，人类的理性是有限的，因此所有的决策都是基于有限理性（Bounded Rationality）的结果。这位天才科学家继而提出，如果能利用存储在计算机里的信息来辅助决策，人类理性的范围将会扩大，决策的质量就能提高。

他进而预测：在后工业时代，也就是信息时代，人类社会面临的中心问题将从如何提高生产率转变为如何更好地利用信息来辅助决策。

西蒙教授毕业于芝加哥大学，1943年获得政治学博士学位，此后半个多世纪，他长期在卡内基梅隆大学任教。

卡内基梅隆大学，是美国信息技术研究的“火车头”，它以计算机科学和“交叉性研究”闻名于世。西蒙的整个学术生涯都浸润着卡内基梅隆的色彩，他从政治、经济出发，把毕生的精力都集中在对决策和信息的研究上，将不同学科之间的“交叉性”应用得炉火纯青，也硕果累累。1975年，由于对人工智能的贡献，他获得了计算机学界的最高奖项：图灵奖；1978年，他又因为对“商务决策过程”的出色研究戴上了诺贝尔经济学奖的桂冠。

追本溯源，学界普遍认为，西蒙对决策支持系统的研究，是现代商

务智能概念最早的源头和起点。[04]但西蒙可能没有想到，他播下的“决策支持”的种子，在半个世纪以后，却结出了“商务智能”的果实，并成为信息时代的一朵奇葩。

从决策支持系统到商务智能，名字变了，但新瓶装的还是旧酒。现代商务智能技术回答的还是决策支持系统面对的老问题：如何将数据、信息转化为知识，扩大人类的理性，辅助决策？

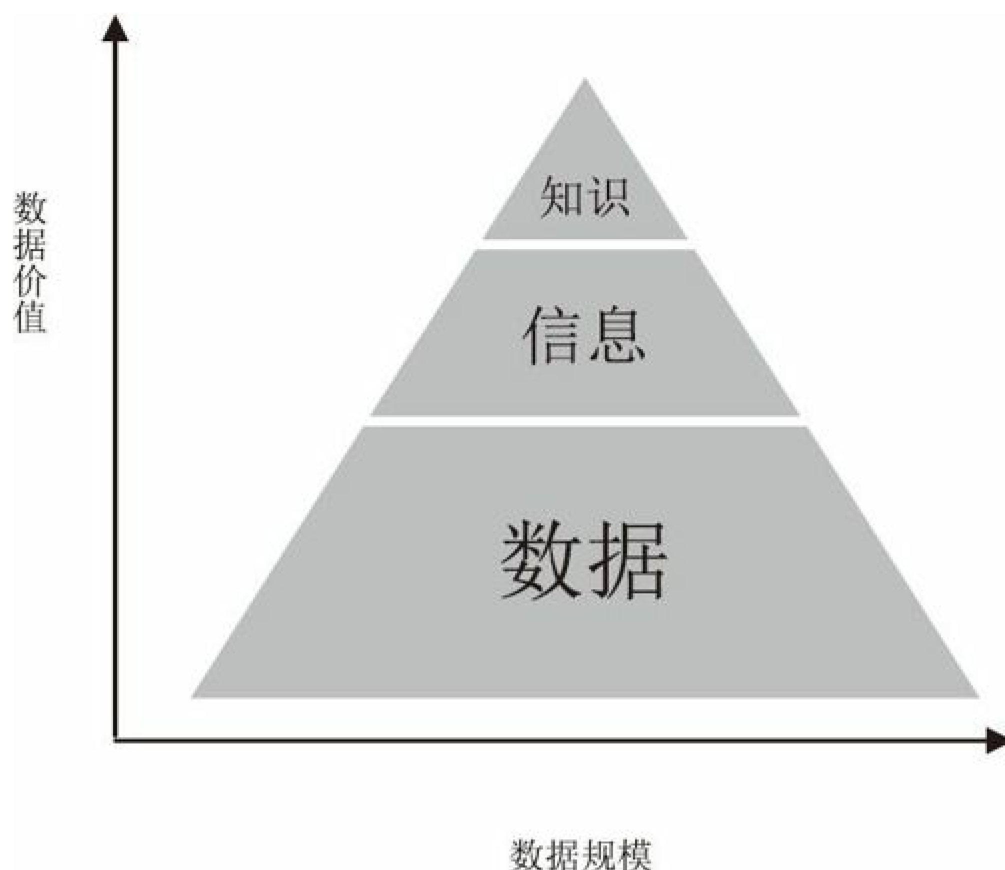
从数据到知识，这个跨越，人类用了半个多世纪。

在半个多世纪的漫长过程中，决策支持系统曾经一度因为缺乏有效的数据组织方式而徘徊不前。直到上世纪90年代，由于若干新技术的出现，打破了瓶颈，“商务智能”的概念才横空出世。随后，其发展取得了前所未有的加速度，在本世纪第一个十年蓬勃向上。今天，回头考察这些新技术的一一出现，可以清楚地看到商务智能的产业链条不断向前延伸的轨迹。

从数据、信息到知识的演变




数据、信息和知识的区别和联系




这个轨迹的起点当然就是计算机。计算机，是硬件和软件相结合的产物。它的发明，是诸多不同领域的科学家共同努力的结果。冯·诺伊曼其实是一名数学家，他之所以被称为“计算机之父”，其最大的贡献之一，在于他明确了计算机内部的数据组织形式：二进制。

二进制的引进，解决了在没有“情感、智能和生命”的物理机器中表达、计算、传送数据的最大难题，有了二进制，软件的运行才有了支点。

如前文所述，软件是由程序和数据组成的。二进制的确定，解决了数据在计算机内部传送“理解”和“流动”的问题，但当数据在计算机内部累积得越来越多的时候，如何快速地组织、存储和读取数据又成为新的

挑战。

计算机科学家一直在研究数据在软件内部的最佳组织方式。1970年，IBM的研究员埃德加·科德（Edgar Codd）发明了关系型数据库，成为软件发展历史上一个跨越性的里程碑。

此前，数据库的组织结构以网状、层级制为主，复杂多变，程序和数据之间你中有我、我中有你，彼此有很强的依赖性。科德提出的关系型数据库具有结构化高、冗余度低、独立性强等优点，彻底把软件中的程序和数据分立开来。从此，软件的发展成了“两条腿”走路，程序和数据在各自的轨道上自由奔跑。

科德后来又总结出构建关系型数据库的“黄金十二定律”，把理论扎扎实实地推向了实践，关系型数据库开始得到大范围地推广，引发了一场软件领域的革命。科德也因此获得1981年的图灵奖。

此后，大型软件，即大型信息管理系统的应用一日千里、遍地开花。

这些信息系统的建立和运行，使人类从繁杂的重复性劳动当中解放出来，大大提高了商业效率。但这些信息系统，都是针对特定的业务过程、处理离散事务的“运营式”信息系统。

所谓“运营式系统”，是指为提高日常工作的效率而设计的系统，数据在其中的作用，是一个个商务流程的记录，数据在这些系统内不断累积的结果，仅仅用于查询，而不是分析。

上个世纪90年代，面对信息管理系统的普及、各行各业数据记录的激增，管理大师彼得·德鲁克（Peter Drucker）曾发出慨叹：迄今为止，

我们的系统产生的还仅仅是数据，而不是信息，更不是知识！[05]

怎样从各个独立的信息系统中提取、整合有价值的数据，从而实现从数据到信息、从信息到知识、从知识到利润的转化？这个要求，随着信息管理系统的普及，变得越来越迫切。企业的规模越来越庞大、组织越来越复杂，市场更加多变、竞争更加激烈，信息是否及时准确、决策是否正确合理，对组织的兴衰存亡影响越来越大，一步走错，可能全盘皆输。

由于实业界这些迫切的需要，决策支持系统的旧问题又重新占据了顶尖科学家的大脑。

商务智能的“幽灵”开始徘徊……

# 结蛹：数据仓库之厚积薄发

岳不群叹了口气，缓缓地道："三十多年前，咱们气宗是少数，剑宗中的师伯、师叔占了大多数。再者，剑宗功夫易于速成，见效极快。大家都练十年，定是剑宗占上风；各练二十年，那便是各擅胜场，难分上下；要到二十年之后，练气宗功夫的才渐渐地越来越强；到得三十年时，练剑宗功夫的便再也不能望气宗之项背了。然而要到二十余年之后，才真正分出高下，这二十余年中双方争斗之烈，可想而知。"

——金庸，《笑傲江湖》第九章，1967年

决策支持系统面临的"瓶颈式"难题，是如何有机地聚集、整合多个不同运营信息系统产生的数据。对这个问题的关注起源于美国计算机科学研究的另一所重镇：麻省理工学院。和卡内基梅隆大学一起，这两所大学先后为现代商务智能的发展奠定了主要的基石。

20世纪70年代，麻省理工学院的研究人员第一次提出，决策支持系统和运营信息系统截然不同，必须分开，这意味着要为前者设计独立的数据存储结构。但受限于当时的数据存储能力，该研究在确立了这一论点后便停滞不前。

但这个研究如灯塔般为实业界指明了方向。1979年，一家以决策支持系统为己任、致力于构建独立数据存储结构的公司Teradata诞生了。Tera，是太字节，其大小为$2^{40}$，Teradata的命名表明了公司处理海量数据的决心。1983年，该公司利用并行处理技术为美国富国银行（Wells Fargo Bank）建立了第一个决策支持系统。这种先发优势令Teradata至今一直雄踞在数据行业的龙头榜首。

另一家信息技术的巨头——国际商业机器公司（IBM）也在为集成企业内不同的运营系统大伤脑筋。越来越多的IBM客户要面对多个分立系统的数据整合问题，这些处理不同事务的系统，由于不同的编码方式和数据结构，像一个个信息孤岛，处于老死不相往来的状态。1988年，为解决企业的数据集成问题，IBM公司的两名研究员（Barry Devlin和Paul Murphy）创造性地提出了一个新的术语：数据仓库（Data Warehouse）。

一声惊雷，似乎宣告了数据仓库的诞生。可惜IBM在首创这个概念之后，也停步不前，只把它当做一个花哨的新名词用于市场宣传，而没有乘胜追击、进一步提出实际的架构和设计。IBM很快在这个领域丧失其领先地位；2008年，IBM甚至通过兼并Cognos才使自己在商务智能的市场上重占一席之地，这是后话。

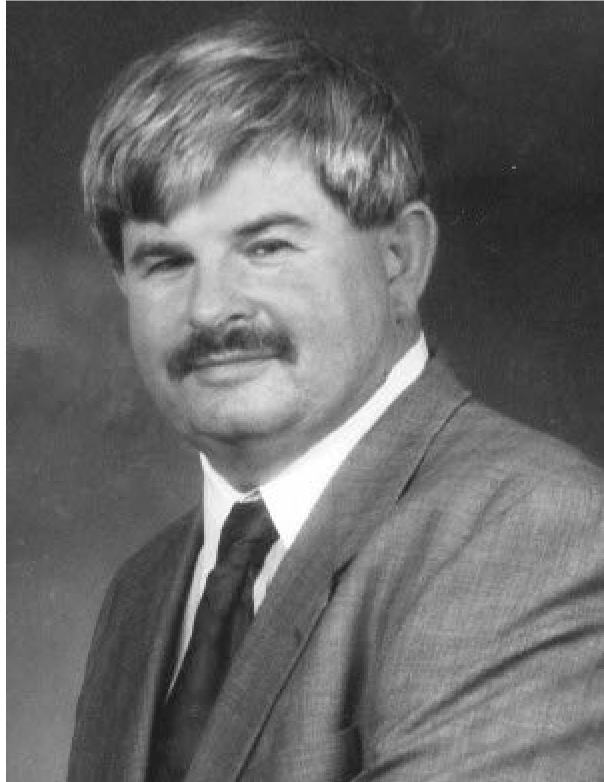
但这之后，更多的信息技术企业垂涎于数据仓库的“第一桶金”，纷纷开始尝试搭建实验性的数据仓库。

又是几年过去，1992年，尘埃终于落定。比尔·恩门（Bill Inmon）出版了《数据仓库之构建》（*Building the Data Warehouse*）一书，第一次给出了数据仓库的清晰定义和操作性很强的实战法则，真正拉开了数据仓库走向大规模应用的序幕。恩门不仅是长期活跃在这个领域的理论领军人物，还是一名企业家。此后，他的“江湖地位”也得以确定，被誉为“数据仓库之父”。

恩门所提出的定义至今仍被广泛地接受：

“数据仓库是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化（Time

Variant）的数据集合，用于支持管理中的决策制定。”



比尔·恩门：数据仓库之父

2007年曾被《计算机世界》评为近40年计算机产业最具影响力的十大人物之一，目前还活跃在数据仓库领域，他的最新成果是将“非结构化的文本数据”通过特定的工具装入数据仓库。

数据仓库和数据库的最大差别在于，前者是以数据分析、决策支持为目的来组织存储数据，而数据库的主要目的则是为运营性系统保存、查询数据。

江山代有才人出。

恩门一统江湖没多久，风头又被拉尔夫·金博尔（Ralph Kimball）抢了去。金博尔是斯坦福大学毕业的博士，长期在决策支持系统的软件公司工作。1996年，他也出版了一本书：《数据仓库的工具》（*The Data Warehouse Toolkit*），金博尔在书里认同了比尔·恩门对于数据仓库的定义，但却在具体的构建方法上和他分庭抗礼。



拉尔夫·金博尔

他的数据仓库构建方法目前在市场上占据了主流。和普适计算的创始人马克·韦泽一样，他也曾经在施乐公司的帕罗奥多研究中心（PARC）长期工作过。（图片来源：datamgmt.com网站）

恩门强调数据的一致性，主张由顶至底的构建方法，一上来，就要

先创建企业级的数据仓库。金博尔却说：不！务实的数据仓库应该从下往上，从部门到企业，并把部门级的数据仓库叫做“数据集市”（Data Mart）。两人针锋相对，各自的追随者也唇舌相向，很快形成了明显对立的两派。

两派的异同，就好比华山剑法的气宗和剑宗。主张练“气”的着眼全面和长远，耗资大，见效慢；主张练“剑”的强调短、平、快，效果可能立竿见影。

如金庸在《笑傲江湖》中描写的剑气之争一样，两派华山论剑的结果不难猜测，金博尔“从易到难”的架构迎合了人类的普遍心理，大受欢迎，商务界随即掀起了一股创建数据集市的狂潮。“吃螃蟹”的结果，有大面积的企业碰壁撞墙，也有不少企业尝到了甜头，赚了个盆满钵盈。

潮起潮落中，两派又有新的融合和纷争。油灯越拨越亮，道理越辩越明，数据仓库的理论和技术，在争论中不断地得以丰富，到2000年，其理念和架构，已经完全成熟，并被业界所接受。

如蚕之蛹，数据仓库是商务智能的依托，是对海量数据进行分析的核心物理构架。它可以形象地理解为一种格式一致的多源数据存储中心，数据源可以来自多个不同的系统，如企业内部的财务系统、客户管理系统、人力资源系统，甚至是企业外部的系统；这些系统，即使运行的平台不同、编制的语言不同、所处的物理位置不同，但其数据可以按统一定义的格式被提取出来，再通过清洗、转换、集成，最后百流归海，加载进入数据仓库。这个提取、转换、装载的主要过程，可以通过专门的ETL（Extraction, Transformation, Load）工具来实现，这种工具，如今已是数据仓库领域的主打产品。

ETL工具和数据仓库理论的成熟，突破了决策支持系统的瓶颈。从此，商务智能的发展走上了顺风顺水的“快车道”，接下来，好戏连台上演。

# 蚕动：联机分析之惊艳

当越来越多的组织认识到联机分析的需要以及其带来的巨大收益的时候，分析型的用户就会增加。在人类的历史上，只有很少一部分运筹学专家曾经负有这样的责任：为企业开展如此高端的分析。[06]

——埃德加·科德，关系型数据库之父，1993年

数据仓库的物理结构出现以后，活跃在前沿的科学家一下子找到了自己的专属“阵地”，商务智能的下一个产业链：联机分析，如水到渠成般迅速形成。数据仓库开始散发真正的魅力。

联机分析（Online Analytical Processing），也称多维分析，本意是把分立的数据库“相联”，进行多维度地分析。

“维”是联机分析的核心概念，指的是人们观察事物、计算数据的特定角度。例如，跨国零售商沃尔玛如果要分析自己的销售量，它可以按时间序列分析、商品门类分析、地区国别分析，也可以按进货渠道分析、客户群体分析，这些不同的分析角度，就叫“维度”。

分析问题的任何角度，都可以视为一个或多个维度的交叉。例如：

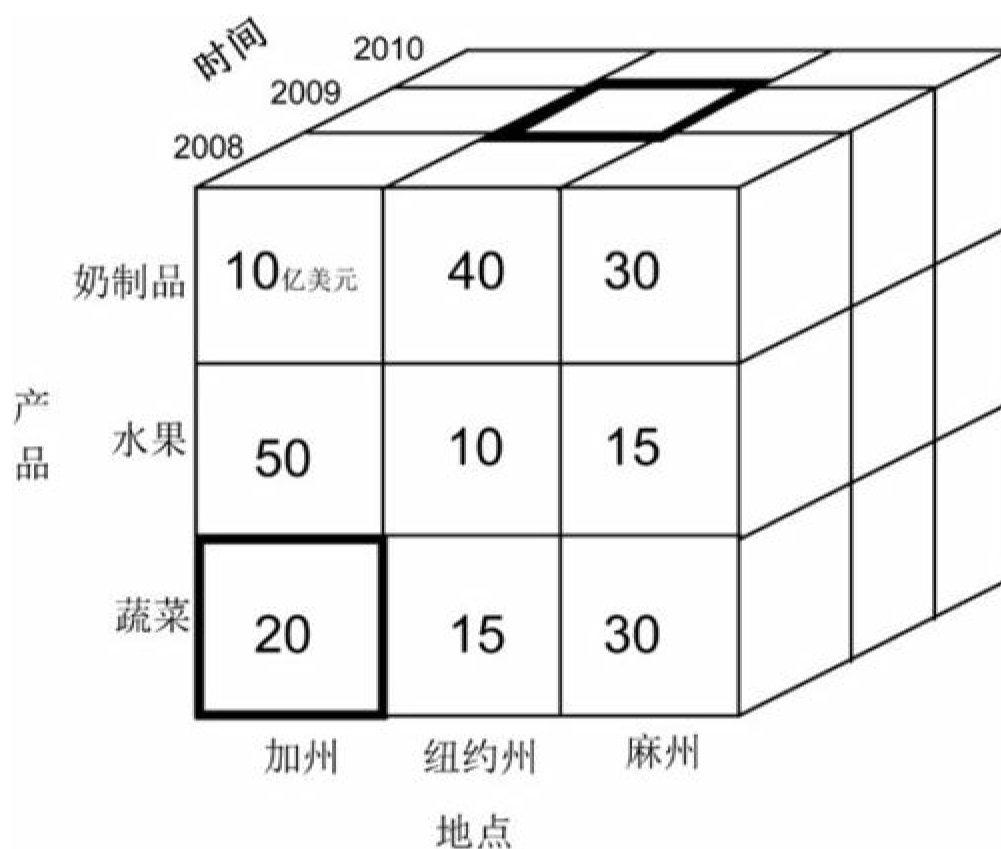
沃尔玛2011年在美国纽约州的销售量是多少？这是个“地区”和“时间”两个维度交叉的问题。

沃尔玛2011年在纽约州奶制品的销售量是多少？这是个“地区”、“时间”和“产品类别”三个维度交叉的问题。

沃尔玛2011年在纽约州进口奶制品的销售量是多少？这是个“地区”、“时间”、“产品类别”及“供货渠道”四个维度交叉的问题。

随着维度的不断增多，问题可能变得很复杂。三个维度就是三度空间，也可以想象成一个立方体。一旦超过了三个维度，人类的思维和想象能力就受到了很大的限制。

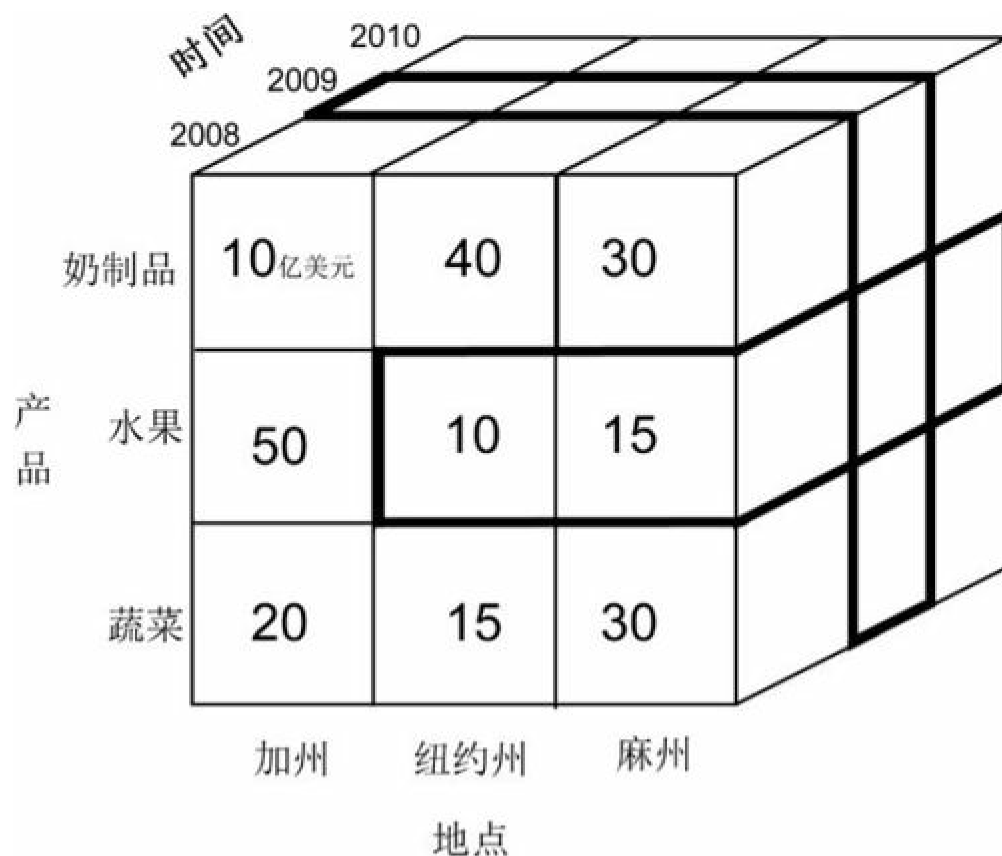
理解一个维度或两个维度的交叉




加粗方块（上）：表示2009年纽约州奶制品的销售量

加粗方块（下）：2008年加州蔬菜的销售量为20亿美元

理解时间、产品和地点三个维度的交叉




横割（地点和产品两个维度的交叉）：所有年份（本图只有2008年至2010年）纽约州和麻州水果的总销量，为横向加粗部分

竖切（时间一个维度）：所有州、所有产品（本图只有3个州、3种产品）在2009年的总销量，为纵向加粗部分

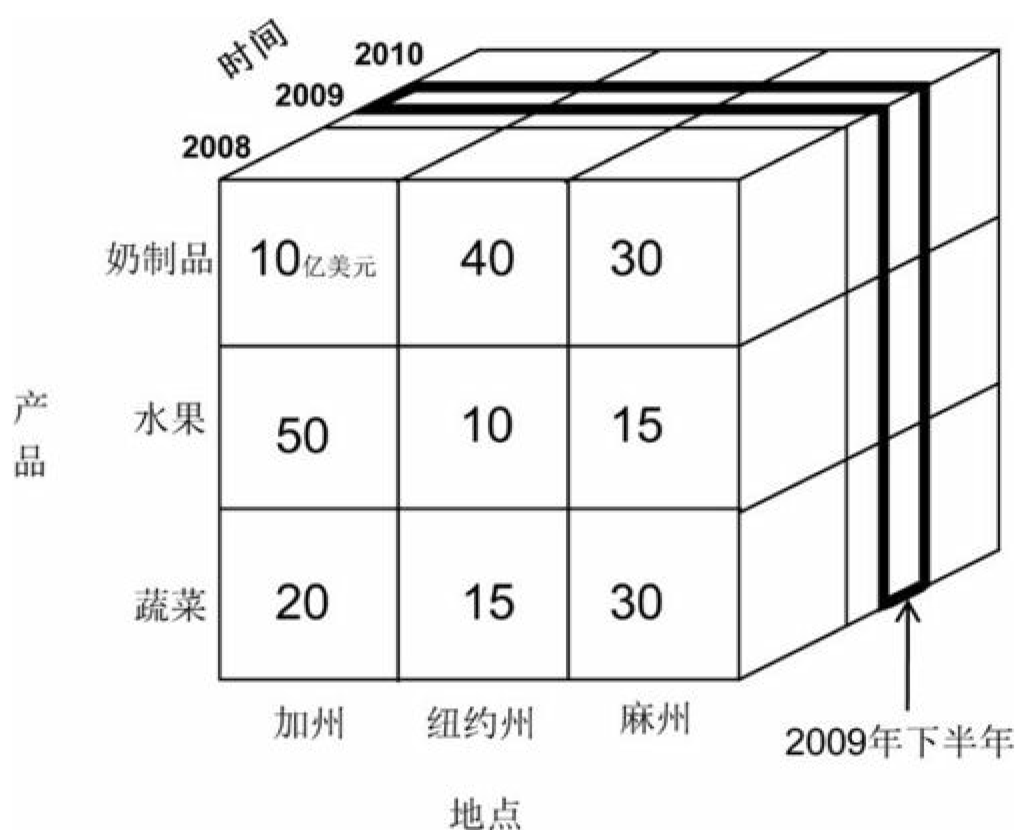
说明：为了绘图的方便，这个例子每一个维度只取了3个值。事实上，每个维度的值都可以无限制增加，例如，年份可以增加2005、2006、2007年等，产品可以增加甜点、咖啡等，地点可以增加佛州等。

同时，一个维度，还可以下钻细分（drill down）。例如，就时间的维度而言，问完了一年的销量，分析人员可能会对半年的、一个季度的

销量，甚至每个月、每一天的销量感兴趣；又如，就地点的维度而言，知道了纽约州的销量，分析人员可能又立刻想知道某个地区、某个城市甚至某个小区的销量。

和下钻相对应的，是上卷（roll up），例如，从各个州的销量，加总到全美国的销量，就是一个典型的上卷。

理解下钻和上卷



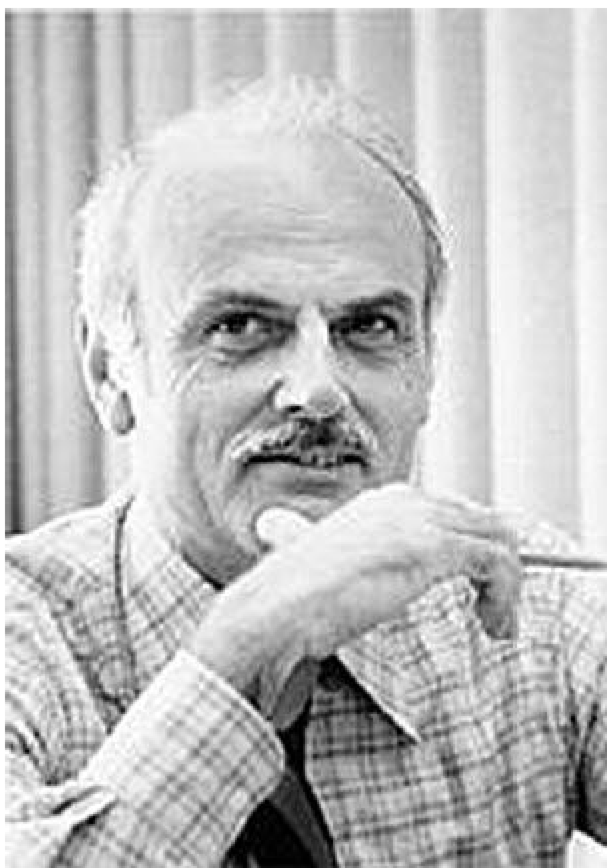

下钻：可以按“年—半年—季度—月—日”的层级一直下钻到每一天
加粗部分：下钻到半年——2009年下半年3个州、3种产品的总销量

上卷：这个例子里只有3个州，如果有50个州，就可以上卷到全国

以关系型数据库为基础的运营式信息系统，事实上，也可以回答以上任何一个问题，但它回答问题的方式，是通过事先设计的报表，也就是说，根据用户指定的条件，由软件开发人员事先一一定制，通过“一对一”的查询，将结果通过报表的形式返回给用户。

报表，是关系型数据库时代将数据转化为信息和知识最主要的手段。

基于一或两个维度的分析，是简单报表；交叉的维度越多，报表就越复杂，而且不同维度的组合将产生不同的报表，对一个立足于决策的用户来说，他的需要是“动态”的：他可能问出任意维度交叉和细分的问题，但软件开发人员只能将最常见的问题定制在软件中。没有定制的问题，系统就无法回答。所以，在联机分析技术出现之前，这种静态的、固定的报表根本无法满足决策分析人员的全部需要。



埃德加·科德（1923-2003）

英国人，1948年移居美国，加盟IBM，因提出关系型数据库，获得1981年图灵奖。1993又率先定义了“联机分析”（OLAP）。（图片来源：维基百科）

早在1960年代，研究人员就意识到了这种“动态”决策需求和“静态”报表之间的矛盾，决策支持系统的先行者就开始探索联机分析的方法。1970年，第一个联机分析的产品就已经问世。它通过建立一个复杂的、中介性的“数据综合引擎”，把分布在不同系统的数据库人为地联结起来，实现了联机分析。

1993年，发明关系型数据库的科德再一次站到了创新的潮头。他发表了论文《信息技术的必然：给分析用户提供联机分析》（Providing OLAP to User-Analysts: An IT Mandate），在这篇文章中，他详尽地阐述了联机分析的定义，并为如何构建联机分析提出了“黄金十二定律”。他形象地比喻说：

“用关系型数据库来分析数据，是试图用‘锤子’把一个‘螺丝钉’硬生生地‘敲’进墙，虽然最后可以勉强完成任务，但很费劲，为什么不用‘螺丝刀’呢？”

科德认为，联机分析就是解决“数据分析”问题的“螺丝刀”。其惊艳之美在于用户可以根据自己的需要随时创建“万维”动态报表，也就是说，报表的定制权由后台的开发人员直接转移到了前端的用户。

有了联机分析，用户可以自己随时创建自己所需要的报表，开发人员只需要预先为用户在后台构建多维的数据立方体（Cube）。一旦多维立方体建模完成，用户可以在前端的各个维度之间自由切换，并可以从不同的维度、不同的粒度对数据进行分析，从而获得全面、动态、可随时加总或细分的分析结果。在多维立方体的构建和运算方面，曾在IBM和微软工作过的詹姆斯·格雷（James Gray）多有贡献，他也于1998年获得了图灵奖。

因为有了数据仓库，不再需要不同数据库之间的人为“联机”，联机分析找到了真正的用武之地，如有源之水，活力四射。任何复杂的报表，都可以在鼠标的瞬间点击之下从用户的指尖弹出，数据尽在手中，如玲珑剔透的水晶体，任意横切竖割，流畅的美感令人叹为观止。

# 破茧：数据挖掘之智能生命的产生

每天早上一醒来，我就要问自己：怎么才能让数据流动得更好、管理得更好、分析得更好？[07]

——罗林·福特，沃尔玛首席信息官

数据仓库、联机分析技术的发展和成熟，为商务智能奠定了框架，但真正给商务智能赋予“智能”生命的是它的下一个产业链：数据挖掘。

一开始，数据挖掘曾一度被称为“基于数据库的知识发现”（Knowledge discovery in database）。随着数据仓库的产生，“数据挖掘”的叫法开始被广泛接受。也正是因为有了数据仓库的依托，数据挖掘如虎添翼，如“巧妇”走进了“米仓”，在实业界不断创造点“数”成金的故事。其中，最为经典的例子当属啤酒和尿布。

这是一个关于零售帝国沃尔玛的故事。

沃尔玛，是全世界最大的零售商，拥有8400多家分店、200多万雇员；它的人数，和美国联邦政府的雇员等量齐观；它的收入，2010年突破了4000亿美元，超过了很多国家的GDP总值。

沃尔玛拥有世界上数一数二的数据仓库，是最早应用数据挖掘技术的企业之一，也是数据挖掘技术的集大成者。在一次例行的数据分析之后，研究人员突然发现：跟尿布一起搭配购买最多的商品竟是啤酒！

数据挖掘（Data Mining）

数据挖掘是指通过特定的计算机算法对大量的数据进行自动分析，从而揭示数据之间隐藏的关系、模式和趋势，为决策者提供新的知识。

之所以称之为“挖掘”，是比喻在海量数据中寻找知识，就像开矿掘金一样困难。

尿布和啤酒，听起来风马牛不相及，但这是对历史数据进行挖掘的结果，反映的是数据层面的规律。

这种关系令人费解，这是一个真正的规律吗？

经过跟踪调查，研究人员终于发现事出有因：一些年轻的爸爸经常要到超市去购买婴儿尿布，有30％到40％的新爸爸会顺便买点啤酒犒劳自己。沃尔玛随后对啤酒和尿布进行了捆绑销售，不出意料，销售量双双增加。

沃尔玛还有很多利用数据挖掘扩大销售的故事。2004年，分析人员发现，每次飓风来临，一种袋装小食品“Pop-Tarts”的销售量都会明显上升。手电筒、电池、水，这些商品的销量会随着飓风的到来而上升，很容易理解，但Pop-Tarts的上升是不是必然的呢？

研究人员后来发现，这也是一个有用的规律：Pop-Tarts的销量上升，一是因为美国人喜欢甜食，二是因为它在停电时吃起来非常方便。此后，飓风来袭之前，沃尔玛也会提高Pop-Tarts的仓储量，以防脱销，并把它和水捆绑起来销售。

如果没有数据挖掘，Pop-Tarts和飓风的微妙关系就难以被发现。

1989年，可谓数据挖掘技术兴起的元年。

这一年，图灵奖的主办单位计算机协会（ACM）下属的知识发现和数据挖掘小组（SIGKDD）举办了第一届数据挖掘的学术年会，出版了专门期刊。此后，数据挖掘一直被热捧，其发展如火如荼，甚至成为一门独立的课目走进了大学课堂；在美国的不少大学，还先后设立了专门的数据挖掘硕士学位。

也正是1989年，高德纳咨询公司的德斯纳（Howard Dresner）在商业界为“商务智能”给出了一个正式的定义：

“商务智能（Business Intelligence），指的是一系列以事实为支持、辅助商业决策的技术和方法。”

这个定义，强调了商务智能是一系列技术的集合，获得了业界的广泛认同。

商务智能的概念在1989年完全破茧而出，并不是历史的巧合，而是因为数据挖掘这种新技术的出现，使商务智能真正具备了“智能”的内涵，也标志着商务智能完整产业链的形成。

如果说联机分析是对数据的一种透视性的探测，数据挖掘则是对数据进行挖山凿矿式的开采。它的主要目的，一是要发现潜藏在数据表面之下的历史规律，二是对未来进行预测，前者称为描述性分析，后者称为预测性分析。沃尔玛发现的啤酒和尿布的销售关联性就是一种典型的描述性分析；考察所有历史数据，以特定的算法对下个月啤酒的销售量进行估测以确定进货量，则是一种预测性分析。

数据挖掘把数据分析的范围从“已知”扩大到了“未知”，从“过去”推

向了“将来”，是商务智能真正的生命力和“灵魂”所在。它的发展和成熟，最终推动了商务智能在各行各业的广泛应用。

数据挖掘的两个侧重点




通过十多年的发展，数据挖掘的范围正在不断扩大。传统的数据挖掘是指在结构化的数据当中发现潜在的关系和规律，但随着商业竞争的白热化，更加高端的数据挖掘也开始初现端倪。例如，通过网络留言挖掘顾客的意见。顾客在博客、论坛、社交网站和微博上用文字记录的消费体验，对商品和服务发表的看法和评价，是一种非结构化的数据。如何把散布在网络上的这些资源整合起来，并从中自动挖掘有价值的信息和知识，正是当前数据挖掘面临的最大挑战之一。数据仓库之父比尔·恩门近年来就在这个领域多有建树。

结构化数据和非结构化数据

按结构，数据可以划分为两类：结构化数据和非结构化数据。

结构化数据是指存储在数据库当中、有统一结构和格式的数据，这种数据，比较容易分析和处理。非结构化数据是指无法用数字或统一的结构来表示的信息，包括各种文档、图像、音频和视频等，这种

数据，没有统一的大小和格式，给分析和挖掘带来了更大的挑战。

从结构化数据到非结构化数据的推进，也代表着可供挖掘的数据在大幅增加。

# 化蝶：数据可视化的华丽上演

图形是解决逻辑问题的视觉方法。[08]

——杰克·伯廷（1918-2010），法国统计学家，1977年

随着数据仓库、联机分析和数据挖掘技术的不断完善，业界曾一度认为，商务智能系统已经功德圆满，很好地完成了智能分析的使命，因此早期商务智能的产业链条只含有这三块。

但技术无止境。

进入21世纪之后，风生水起，新的技术浪潮又使商务智能的产业链条向前延伸了一大步：数据可视化。

> 数据可视化（Data Visualization）
>
> 数据可视化是指以图形、图像、地图、动画等更为生动、易为理解的方式来展现数据的大小，诠释数据之间的关系和发展的趋势，以期更好地理解、使用数据分析的结果。

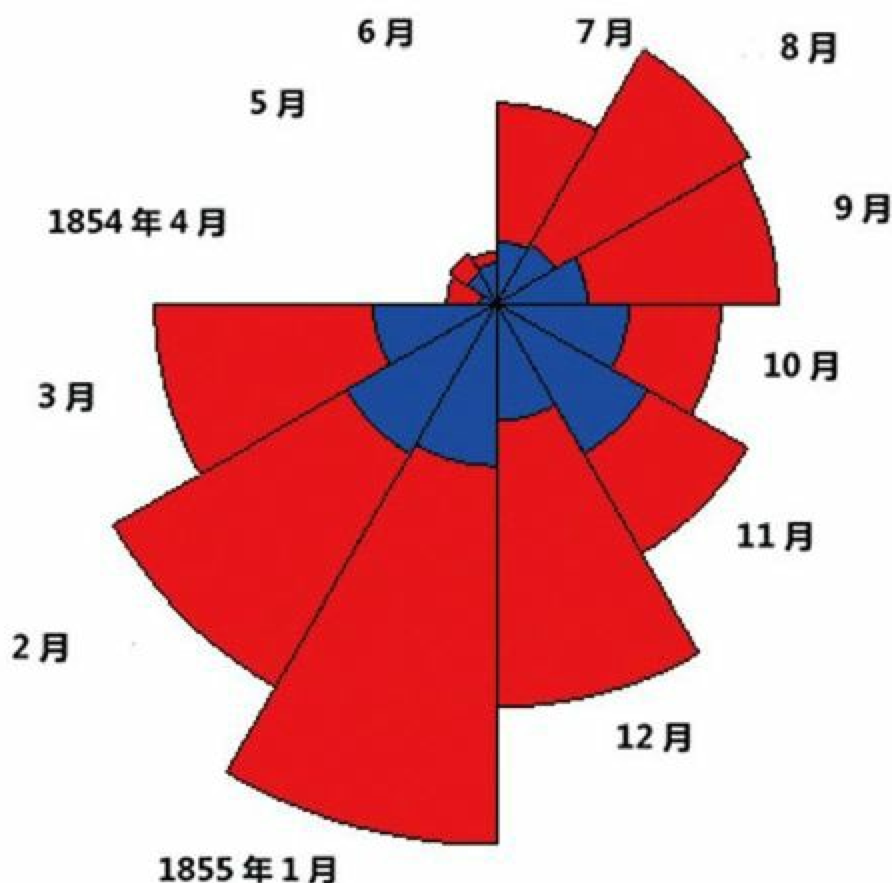
数据可视化也是几代统计学家上百年的梦想。

故事可以追溯到19世纪中期。1850年代，土耳其、英、法等国与俄罗斯之间爆发了克里米亚战争。这场战争共死亡50多万人，异常惨烈。弗罗伦斯·南丁格尔（Florence Nightingale）是英国的一名战地护士，也是一名自学成才的统计专家。她在考察了英国士兵的死亡情况之后，发

现由于医疗卫生条件恶劣导致的死亡人数，大大超出了前线的直接阵亡人数。南丁格尔将她的统计结果制成一个图表，该图表清晰地反映了“战斗死亡”和“非战斗死亡”两种原因死亡人数的悬殊对比，强烈的视觉效果引起了英国社会的极大反响，最后直接促成了英国政府出台建立野战医院的决定。

南丁格尔后来被誉为现代护理业之母，她的这份图形，是历史上第一份“极区图”（Polar Area Diagram），也是统计学家对利用图形来展示数据进行的早期探索。

**1854**年**4**月至**1855**年**3**月，英国军队士兵的死亡原因




图形说明：每月的死亡人数以30°的扇形面积表示，内环蓝色代表因战斗死亡的人数，外环红色代表非战斗死亡的人数，也就是可以预防、改善的医疗卫生原因。（图片来源：SAS公司）

一份图表催生了一座医院，改变了一个制度。

南丁格尔的贡献，充分证明了数据可视化的价值，特别是在公共领域的价值。官僚们麻木的神经尤其需要强烈的视觉效果来冲击、来刺激。生理学也证明，人的大脑皮层当中，有40％是视觉反应区，人类的

神经系统天生就对图像化的信息最为敏感。通过图像，信息的表达和传递将更加直观、快捷、有效。

更重要的原因在于：人的创造力不仅仅取决于逻辑思维，还取决于形象思维。数据可视化的技术，可以通过图像在逻辑思维的基础上进一步激发人的形象思维和空间想象能力，吸引、帮助用户洞察数据之间隐藏的关系和规律。

到了20世纪70年代，由于计算机技术的兴起，美国一批有远见卓识的学者都看到了这个领域巨大的潜力。耶鲁大学的统计学教授弗朗西斯·安斯科姆（F. J. Anscombe）就是其中的先驱人物。1973年，他发表论文《统计分析中的图形》，专门阐述了图形在统计研究当中不可替代的作用。他认为：

“未来的计算机不仅要能计算，还要能将计算结果转变为直观的图形。我们应该研究这两种结果，因为每一种都有助于我们理解问题。”[09]

在这篇文章中，安斯科姆教授提出了“安斯科姆四重奏”，通过这个例子，他强调：在研究数据、使用数据的时候，图形和计算同等重要，有的时候，图形甚至是解决逻辑问题更为直接有效的方法。

这个著名的“四重奏”，是4组同时呈现在你面前的数据（X，Y）。

当你粗略浏览这4组数据之后，你会感觉其数值大多在5到11之间，比较杂乱。稍做对比，你会发现：

X1=X2=X3

X4的值，除一个之外，全部都等于8

Y1≠Y2≠Y3≠Y4

如果再进行简单的统计学计算，很容易得到以下结果：

X1、X2、X3和X4的平均值都等于9，其方差等于10

Y1、Y2、Y3和Y4的平均值都等于7.50，其方差等于3.75

4组数据都符合线性回归：y=3+0.5x

第一组数据

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 10.0 | 8.0 | 13.0 | 9.0 | 11.0 | 14.0 | 6.0 | 4.0 | 12.0 | 7.0 | 5.0 |
| Y1 | 8.04 | 6.95 | 7.58 | 8.81 | 8.33 | 9.96 | 7.24 | 4.26 | 10.84 | 4.82 | 5.68 |

第二组数据

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X2 | 10.0 | 8.0 | 13.0 | 9.0 | 11.0 | 14.0 | 6.0 | 4.0 | 12.0 | 7.0 | 5.0 |
| Y2 | 9.14 | 8.14 | 8.74 | 8.77 | 9.26 | 8.10 | 6.13 | 3.10 | 9.13 | 7.26 | 4.74 |

第三组数据

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X3 | 10.0 | 8.0 | 13.0 | 9.0 | 11.0 | 14.0 | 6.0 | 4.0 | 12.0 | 7.0 | 5.0 |
| Y3 | 7.46 | 6.77 | 12.74 | 7.11 | 7.81 | 8.84 | 6.08 | 5.39 | 8.15 | 6.42 | 5.73 |

第四组数据

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| X4 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 | 19.0 | 8.0 | 8.0 | 8.0 |
| Y4 | 6.58 | 5.76 | 7.71 | 8.84 | 8.47 | 7.04 | 5.25 | 12.50 | 5.56 | 7.91 | 6.89 |

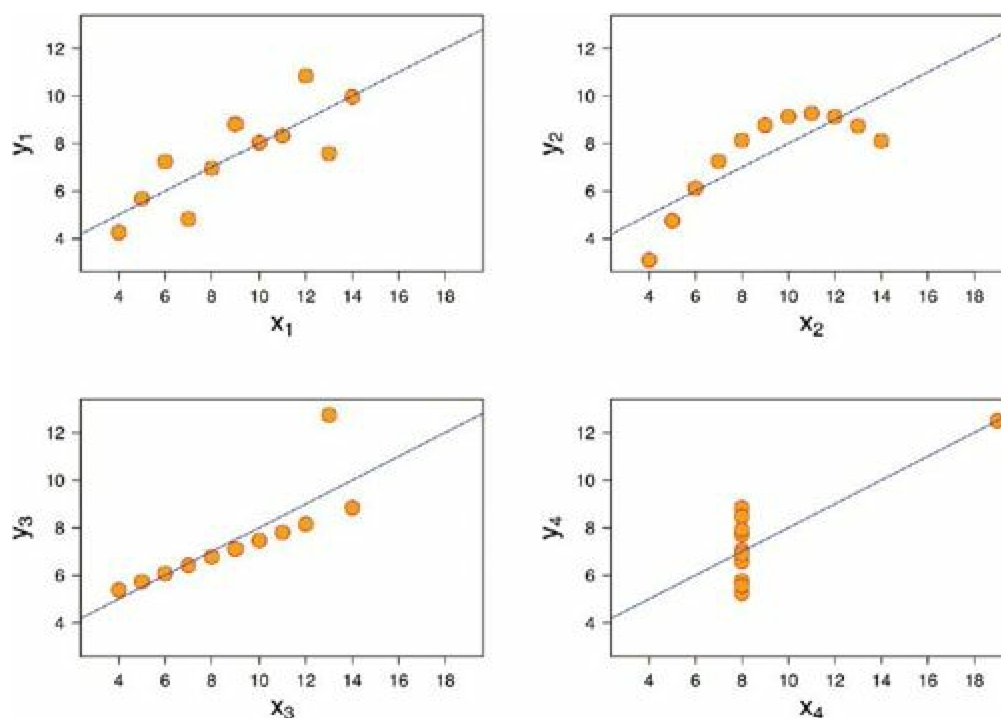
也就是说：

4组数据当中，X和Y之间的关系都是相同的，个别数据的偏离，可以视为随机产生的干扰。

但当我们用散点图把它们在坐标中标出来之后，面对图形，就会立刻发现，统计学“欺骗”了我们：

4组数据当中，仅仅只有第一组数据严格符合利用统计学作出的线性回归结论；

图形是解决逻辑问题的视觉方法：安斯科姆四重奏的真实分布




（图片来源：维基百科）

第二组数据存在某种规律，但显然不是线性的；

第三组数据大部分符合线性回归的模型，但有一对数据明显异常，它是第三对数据（13.0，12.74）；

第四组数据则呈垂直分布，其之所以貌似符合线性回归的分布，是因为其第8对数据（19.0，12.50）在其中起了很大的扭曲作用。

1983年，耶鲁大学的政治学教授爱德华·塔夫特（Edward Tufte）率先奠基了数据可视化这门学科。塔夫特系统地考证了人类用“图形”表达“数据”和“思想”的渊源，整理了种种历史古籍中的图形瑰宝，并结合计算机的发展给统计领域带来的革命，出版了《定量信息的视觉展示》（*The Visual Display of Quantitative Information*）一书。这本书后来被公认为“数据可视化”作为一门学科的开山之作。

这本书的出版，也有一段曲折。因为塔夫特整理了从古到今很多优秀的图表，他坚持要在新书中使用高质量、高精度的彩色插图，几乎所有的出版商都认为这是赔本买卖，没人愿意出版。塔夫特最后无计可施，用自己的房子做了抵押，自费出版了这本书。

结果当然令出版商大跌眼镜：这本书最终获得了很大的商业成功，塔夫特教授也由“政治学”专家成功转型为“信息学”专家。近十多年来，他又先后出版了《视觉解释》（*Visual Explanations*）、《美丽的证据》（*Beautiful Evidence*）等几本重量级的著作，本本都洛阳纸贵，造成了不小的轰动。他本人也成了数据可视化领域当仁不让的掌门人。2010年3月，奥巴马任命塔夫特为顾问，要求他运用“数据可视化”的技术推进联邦政府专项资金使用情况的透明度。

塔夫特教授强调数据可视化的关键在于“设计”，他认为：

“信息过载这回事并不存在，问题出在糟糕的设计，如果你用来表

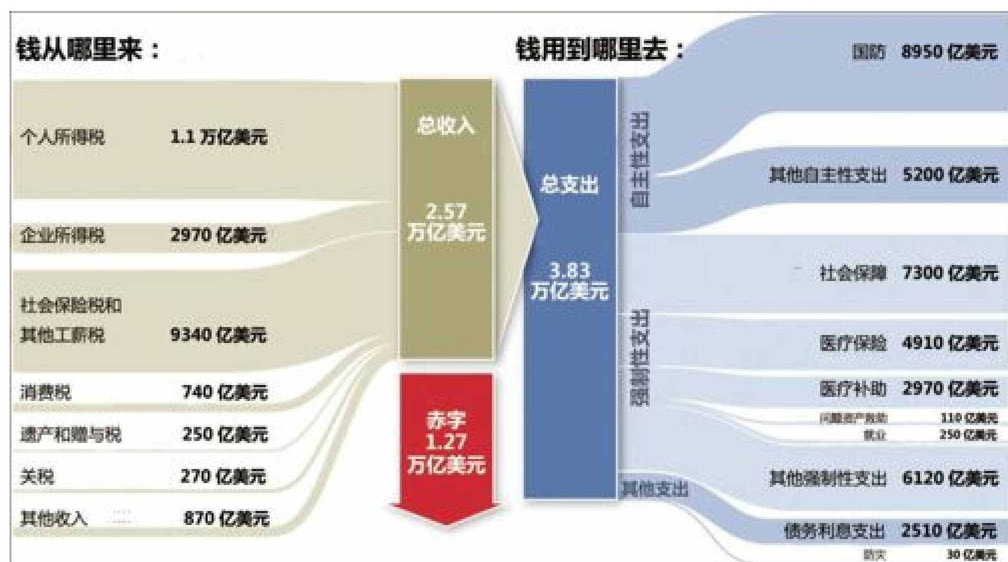
达数据的图形让人感觉杂乱不解，那就要修改你的设计。”[10]

进入21世纪之后，大数据的爆炸使人们急需展示数据、理解数据、演绎数据的工具。这种需求，刺激了数据可视化专业市场的形成，其产品迅速增多，使现在的市场可谓绚丽多彩、百花齐放。从最早的点线图、直方图、饼图、网状图等简单图表，发展到以监控商务绩效为主的仪表盘（dashboard）、记分板（scorecard），到以交互式的三维地图、动态模拟、动画技术等等更加直觉化、趣味化的表现方法，曾经冰冷坚硬、枯燥乏味的数据开始“动”了起来、“舞”了起来，变得“性感”！

数据可视化把美学的元素带进了商务智能，给它锦上添花。一幅好的数据图像不仅能有效地传达数据背后的知识和思想，而且华美精致，如一只只振动翅膀的彩蝶，刺激视觉神经，调动美学意识，令人过目不忘，留下栩栩如生的印象。

2010年2月，奥巴马宣布了联邦政府新的年度预算。《华盛顿邮报》立即对这份新鲜出炉的预算进行了分析报道，它利用数据可视化的技术，抓住了读者的眼球。在图形中以线条的粗细表明各项收支金额的大小，左边是收，右边是支，中间的红色部分是赤字缺口，形象贴切。奥巴马政府收入多少钱，要办哪些事，各项收入与支出的轻重大小，一目了然。

奥巴马政府**2011**年度预算开支的可视化展示



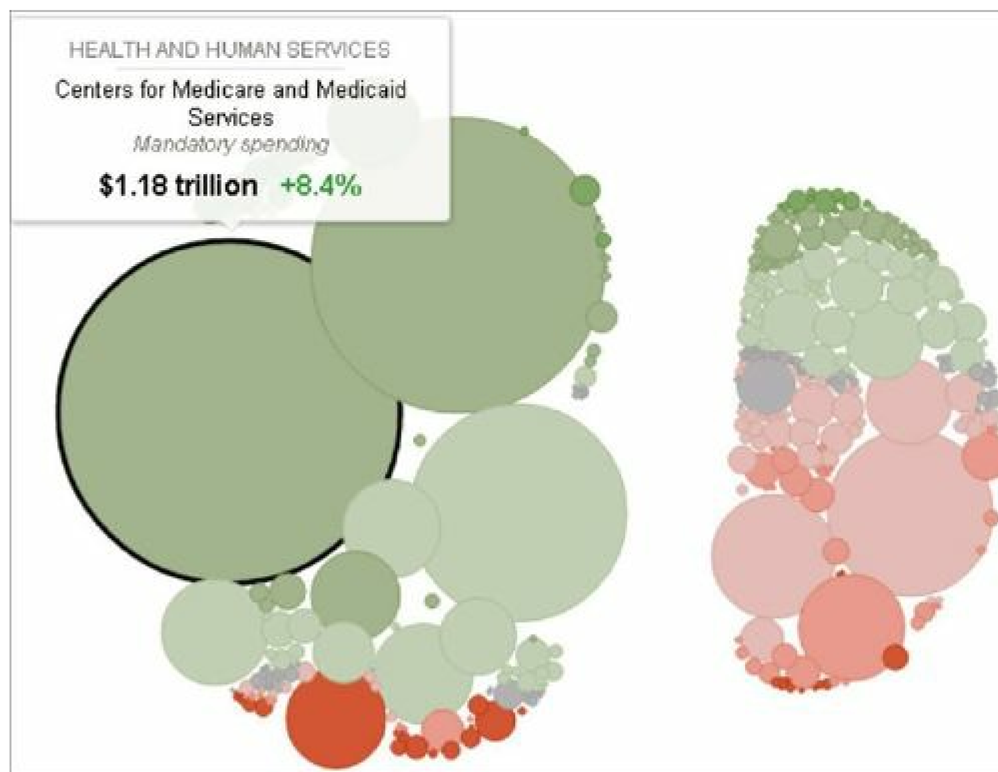

说明：一眼就可以看出，赤字约占美国总支出的1/3，个人所得税是美国政府最大的财政来源，而国防支出是其最大的支出。（图片来源：《华盛顿邮报》，2010年2月1日[11]）

2012年2月，《纽约时报》又用另外一种形式对2013年联邦政府的预算进行了可视化展示。他们用圆形的大小表示金额的多少，颜色表示增减，绿色代表增加，红色代表缩减，变化额度越大，则颜色越深，而且整个图形是动态的，会放大、缩小、移动，也引起了很多读者的兴趣和转载。[12]

每年的10月，诺贝尔奖花落谁家是全世界的热门话题。2011年10月，《福布斯》（*Forbes*）对100多年来各项诺贝尔奖的获得情况做了一个可视化的展示。这是一个以时间为横坐标、以大奖得主的国籍为纵坐标的散点图。不难看出，1940年以前，德国是世界科学和文化的中心，但二次世界大战之后，这个中心毫无疑问转移到了美国。还能看到，美国人的崛起首先在物理领域，其次是医学领域，再次是经济学领域。1969年，开始设立了诺贝尔经济学奖。这之后，美国人几乎囊括了

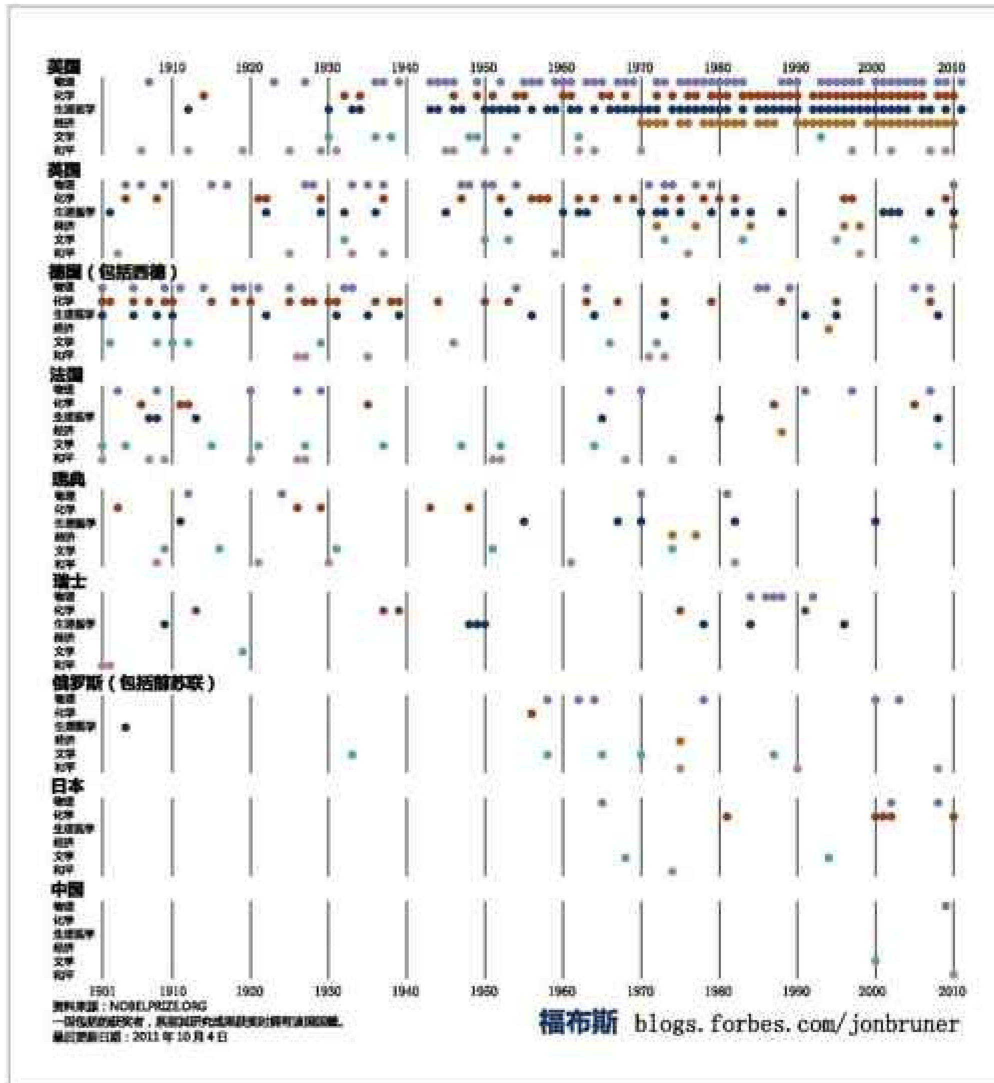
全部的诺贝尔经济学奖。

**2013**年度联邦政府预算开支的可视化展示




说明：左边为强制性开支，右边为自主性开支。强制开支中最大的圆为医疗保险和医疗补助，其为绿色，表示较去年增加了，鼠标停留处显示其大小为1.18万亿，较去年增加了8.4%，是强制性开支中最大的一块。图为网站截屏。

百年诺贝尔奖得主的分布（按国别和奖项）




说明：获奖人的国籍，有时候难以甄别，例如，历史上曾出现以难民身份获奖者。又比如，2009年，高锟获物理学奖时，为英国国籍，但持有香港身份证并居住在香港，制作者将他归入中国。制作者还指出：在美国的314位获得者中，有102位（32％）是在美国本土之外出生的，其中有德裔15位、加拿大裔12位、英国裔10位、俄裔6位、华裔6位；而德国的65位获奖者中，只有11位出生在海外；日本的获奖者，则全是在本土出生的。（图片来源：《福布斯》，2011年10月5日）[13]

作为一个新兴的行业，数据可视化的发展潜力不容小觑。2010年起，谷歌的首席经济学家范里安（Hal Varian）就一直在多种场合强调，下一个十年，将出现一类新的专业人才：数据科学家。其中一种，正是数据可视化工程师，这种人才既懂得数据分析，又精通构图的艺术，集故事讲述和艺术家的特质于一身，将是我们大数据时代的导航员。

数据可视化的这种“导航”作用也极大地推动了商务智能的大众化。通过把复杂的数据转化为直观的图形，并呈现给最普通的用户，商务智能已经不再是少部分高级分析人员的专利，而是贴近大众生活、浅显易懂、人皆可用的工具和手段。

美国联邦政府也意识到“数据可视化”的战略意义。2004年，联邦政府在国土安全部成立了国家可视化分析中心（NVAC），专门推动该项技术在政府部门的应用，特别是在情报分析领域的应用。

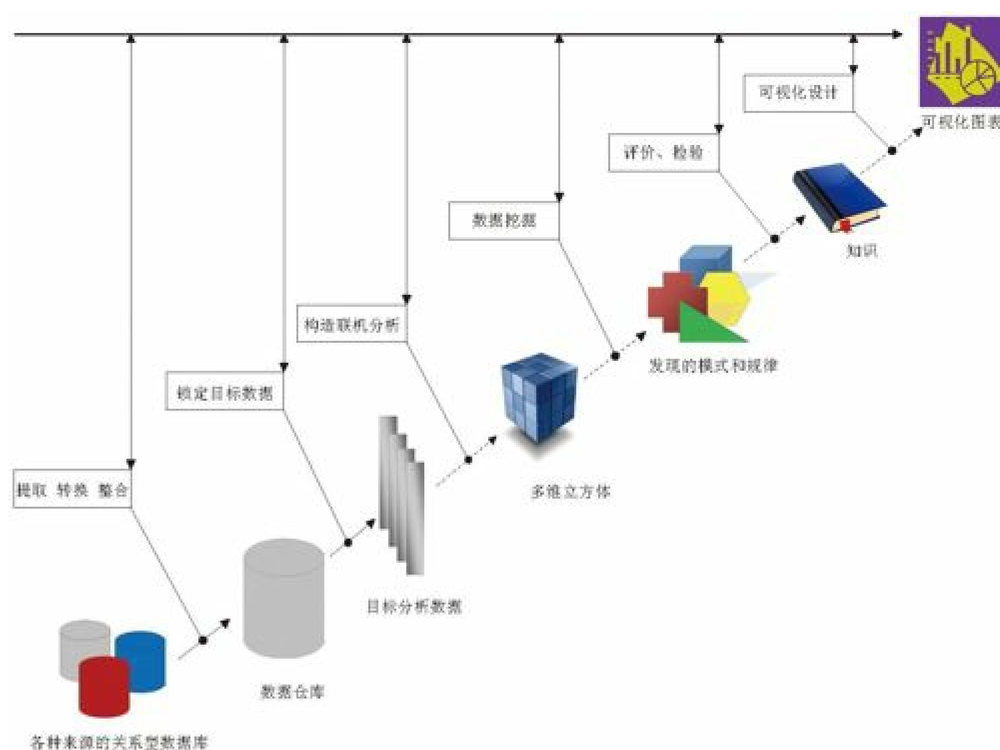
可视化技术的出现，使商务智能的产业链形成了一个从数据整合、分析、挖掘到展示的完整闭环。它的起点是多个独立的关系型数据库，经过数据整合之后形成统一的、多源的数据仓库，再根据用户的需要，重新取出若干数据子集，或构造多维立方体（Cube）进行联机分析，或进行数据挖掘，发现潜藏的规律和趋势。如果挖掘的结果经得起现实的检验，那就形成了新的知识，这种知识，还可以通过数据可视化来表达、展示和传递。

商务智能的这四个产业链，每一块都相当复杂，彼此的独立性也很强。一个好的商务智能产品，并不见得一定要面面俱到，时下不少公司，都专注在一个链条上大做文章。

商务智能的历史，是一个渐进的、复杂的演进过程。至今为止，它的内涵和外延，还处于动态的发展之中。它的各个产业链条，还有不断丰富扩大的趋势。特别是作为其“智能灵魂”的数据挖掘技术，潜力非常巨大，可以预见，将对人类社会的发展产生深远的影响。

大数据时代的竞争，将是知识生产率的竞争。以发现新知识为使命的商务智能，无疑是这个时代最为瞩目的竞争利器。

完整的商务智能流程




说明：虚线代表可选择路径

## 注释

01 A Different Game: Information is Transforming Traditional Businesses, *Economist*, Feb 25th, 2010.

02 英语原文为：“What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention…The scarce resource is not

information; it is processing capacity to attend to information. Attention is the chief bottleneck in organizational activity."—Designing Organizations for an Information-Rich World, Simon, 1971

03 人类第一台计算机到底在哪年发明的，近年来学界对此有所争议。本书以1946年2月在宾夕法尼亚大学发明的ENIAC计算机为准。

04 数据分析技术，在商业领域多被称为"商务智能"，在政府领域则多被称为决策支持系统。

05 英语原文为："Our systems are great at producing data but not at producing information. In our daily life, we deal with huge amount of data and information. Data and information is not knowledge until we know how to dig the value out of it."—*Forbes*, 24 August 1998

06 英语原文为："As more and more organizations recognize the need and significant benefit of OLAP, the number of user analysts will increase. Historically, a small number of experts in operations research have been responsible for performing this type of sophisticated analysis for business enterprises."—Providing OLAP to User-Analysts: An IT Mandate E.F. Codd, 1993

07 英语原文为："Every day I wake up and ask,'How can flow the data better, manage data better, analyze data better?'"—Rollin Ford, Chief Information Officer of Wal-Mart

08 英语原文为："Graphic is the visual means of resolving logical problems."—Graphics and graphic information processing, P.16, Jacques Bertins, 1977

09 英语原文为："A computer should make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."—Graphs in Statistical Analysis, F.J. Anscombe, 1973

10 英语原文为："There is no such thing as information overload, just bad design. If something is cluttered and/or confusing, fix your design."—Edward Tufte

11 Taking apart the federal budget，Data Source: White House Office of Management and Budget; GRAPHIC: Wilson Andrews, Jacqueline Kazil, Laura Stanton, Karen Yourish. *The Washington Post*, Feb1, 2010. 图片翻译处理：肖准。

12 Four Ways to Slice Obama's 2013 Budget Proposal, February 12, 2012, *New York Time*

13 American Leadership in Science, Measured in Nobel Prizes（Infographic）, Jon Bruner, *Forbes*, Oct 5th, 2011. 图片翻译处理：肖准。

# 中　篇　法则博弈
# 大数据时代的发展与危机

在（美国）这个文化传统中，事实是推论的基础，所谓良知，首先是诚实面对全部事实、从真实的前提出发，不在思辨中途偷换概念，然后，让逻辑自己走，它自然会推出结论来。……在重谋略与诡辩的文化中，政治家的聪明才智在拼搏中相互抵消。[01]

——林达，美籍华人作家，2011年

注释

01 《杰斐逊墓碑上忽略了什么》，林达，2011年10月17日。

# 第五章　帝国的法则

一个公众的政府，如果缺乏公众的信息，或者缺乏收集信息的方法和手段，那它只不过是一个短暂的序幕，拉开的将是一场闹剧或悲剧，也可能既是闹剧又是悲剧。知识将永远统治无知：人民如果想要当家做主、主导自己的命运，就必须用知识的力量来武装自己。[01]

——詹姆斯·麦迪逊，第4任美国总统，1822年

联邦政府这个庞大的数据帝国，是如何运作和管理的呢？

没有规矩，难成方圆。这又不得不谈到美国的立法工作。美国是个法制极其发达的国家，上个世纪以来，美国国会、政府先后通过了上百个大大小小的法规，分别对数据的收集、发布、使用和管理等关键环节都做出了具体的规定。

更重要的是，美国人民对于这些法规中不合理的地方，也没有等闲视之，而是不断地提起法律诉讼，最高法院再通过新的判例查缺补漏，推翻国会和政府制定的一些不合理规定。经过几十年的发展，联邦政府对于数据的收集、发布、使用和管理，已经形成了一个比较成熟的框架和体系。

## 收集法则：减负，为人民减负

当收集信息时，每一个机构都要告诉其收集对象以下信息：收集这份信息的目的、将来使用的方式、对其负担的估计，以及这次信息收集工作是志愿的、法定的或是有报酬的。最后，还要告诉他们：如果没有“信息收集许可号”，任何人都可以拒绝填报。

——《纸面工作精减法》第3506（B）（iii）款，1996年

1980年，为了规范、控制联邦政府的信息收集工作，美国国会通过了《纸面工作精减法》（Paperwork Reduction Act）。

该法规定，所有的联邦政府机构，必须在得到行政管理预算局（OMB）批准之后，才能向社会和公众收集数据和信息，国会并且决定，在行政管理预算局新设一个“信息和管制办公室”（OIRA），统管联邦政府的信息收集工作。该办公室还负责发布行业管制标准的职能，这些标准，其实也是数据。关于这个职能，本书将在下一章具体介绍。

任何一个部门，如果要向社会收集数据和信息，必须事先向OIRA办公室提出申请、填报83-I表格，并提前60天在《联邦公告》（Federal Register）上昭告社会，就收集信息的内容和方式听取其他部门和全社会的意见。

走完这个程序，证明了其收集内容的合理性、收集方式的科学性，才能得到OIRA办公室的批准：收到一个由该办公室统一编发的“信息收集许可号”。缺少这个号码，就没有合法性，任何机关都不能向公众收集信息；对没有标明许可号的信息收集表格，公民有权拒绝填报。

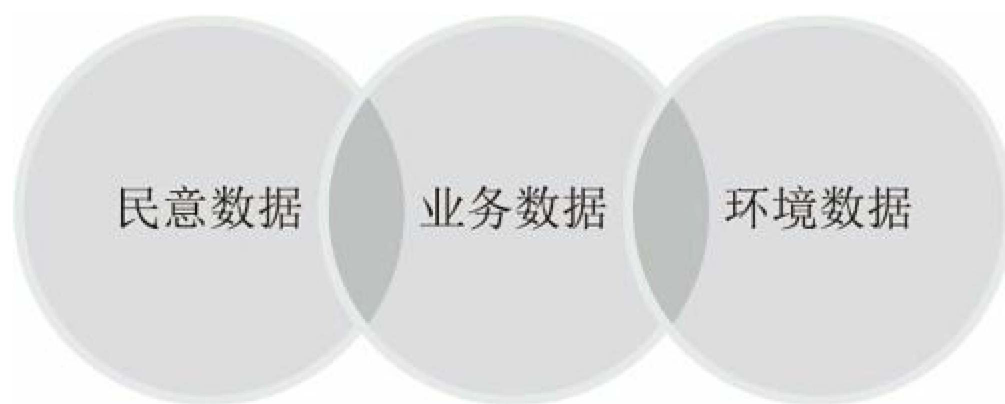
《纸面工作精减法》是规范指导美国联邦政府日常运作的一个重要法规，要理解它的精髓，首先要明白，政府的信息收集工作，需要公民和社会的配合，对公民和社会而言，这是一个负担。1980年国会设立该法的根本目的，就是为了控制这个社会负担、减少“信息扰民”。

本书第三章谈到，联邦政府的数据主要有3个主要来源：一是业务工作的数据，二是社情民意的数据，三是物理环境的数据。这三种数据，互有交叉，各种数据的收集方式，也各不相同。

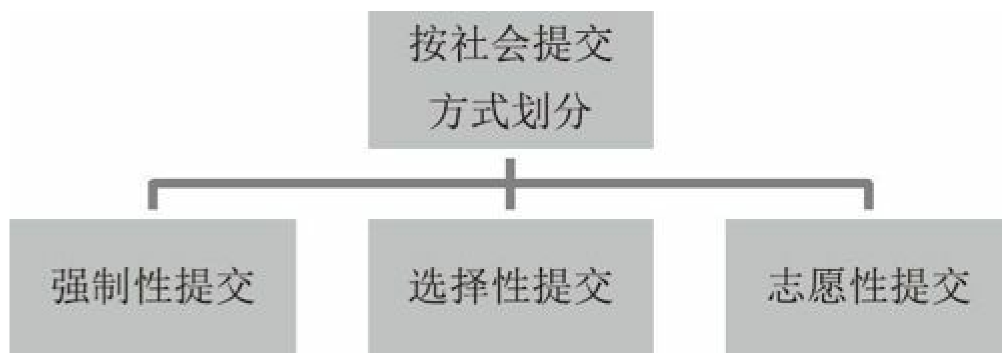
这种划分方法，依据的是数据的性质和来源。

当我们知道联邦政府的信息收集工作对社会是一个负担的时候，可以换一个角度，从法律的角度来考察公民和社会向政府提交信息的义务，就这个角度而言，联邦政府要收集的信息可以划分为强制提交、选择提交和志愿提交3种。

联邦政府数据来源两种不同的划分方法




从数据收集方的角度按数据的性质划分，有交叉




按信息提交方的提交方式划分，无交叉

按提交方式的信息分类

强制提交的信息：公民和社会组织必须依法提交的信息，如企业和个人的纳税情况、工厂的排污量、车间的卫生指标、产品中各种化学物质的含量等等，提交这些信息，是法定的义务。

选择提交的信息：对于一些政府补贴或社会福利项目，公民和组织如果决定申请，就必须填报表格、提交信息，以证实其符合该项目的要求。否则，政府可以拒绝其申请。

志愿提交的信息：多指政府开展的一些基于社情民意的调查，如政府机关为征求公众意见开展的调查，参不参加调查、提不提交数据，完全取决于公民和组织的自愿。正因为公民有权拒绝，联邦政府常常为这些调查提供一些报酬和奖励，以鼓励社会的参与。

信息和管制办公室（OIRA）每年都要编制“年度信息收集预算”（Information Collection Budget），这个“预算”，并不是指联邦政府为了完成信息收集的任务，要花多少钱，或者要投入多少人力和时间，也不是指要收集信息的数量，而是去计算政府机构的信息收集计划会给

全社会带来多大的负担（Information Collection Burden）。这个负担，以小时数来计算。也就是说，美国人民要完成政府规定的信息提交任务，需要投入多少时间。

其计算方法别具一格，有必要简单介绍。

例如，为了完成某项业务工作，某联邦政府机构确定了必须从社会收集某组信息，该收集工作一年需要进行两次，每次需要1000个单位或个人填报，完成一次填报需要一个人投入5个小时，这5个小时包括用于阅读说明、填报数据、检查提交的全部时间。那么，这份信息收集工作的社会负担将是：




根据这个计算方法，该部门将确定这项信息收集工作的年度负担为1万小时，再提交OIRA办公室审核。OIRA办公室如果确定其收集的内容和其他部门没有交叉，收集方法科学合理，将批准该机构1万小时的年度信息收集预算。

这也正是该法被命名为《纸面工作精减法》的原因：通过OIRA办公室把关，科学规划，避免重复收集，确保收集信息的方式简洁有效，尽量减少普通公民和社会组织的信息填报负担。

1995年，联邦政府信息收集负担为65亿小时。

65亿小时是个什么概念呢？

这相当于320万人一年的工作量。也就是说，除了填报表格，320万

人什么也不做，需要一年才能完成联邦政府下达的信息收集任务。

为了应对越来越繁重的信息收集任务，美国国会曾先后于1995年、1996年对《纸面工作精减法》进行了修订。OIRA办公室也曾经制订过“减负计划”：从1996年开始，每年的信息“扰”民时间必须减少10％。

但OIRA办公室明显错误地估计了形势。

这时候的美国，已经不折不扣是一个信息社会，对信息的需求已非人力所能控制。OIRA办公室纵然有心减负、大力控制，也没能挡住各个部门年年递增的信息收集预算。

和1995年相比，1996年联邦政府的信息收集负担不降反升，达68亿小时，其后年年上升。根据OMB每年公布的信息收集预算报告，2009年，联邦政府的信息收集负担达到99亿小时，较10年前的72亿小时上升了37％。

当然，这也从另外一个方面证实了联邦政府所拥有的数据总量确实在大幅上升。

信息收集预算不断攀升的原因主要有两个。一是美国国会在不断制定新的法规，新的法规往往要求收集新的信息，各联邦部门“不得不从”，例如2009年的财政年度，仅仅因为国会2008年通过的新法规，联邦政府的信息收集预算就增加了5500万个小时；[02]二是因为一些选择性提交的数据有很大的波动性，例如，失业人数的增加，会导致申请社会福利人数的增加，从而导致全社会填报信息的负担增加。

**1999—2009**年联邦政府信息收集负担预算明细分析

| 1999—2009年联邦政府信息收集负担预算一览表 | | | |
|---|---|---|---|
| 单位：百万小时 | | | |
| 财政年度 | 联邦政府财政部 | 其他部门机构 | 总计 |
| 1999 | 5,909.74 | 1,274.75 | 7,183.82 |
| 2000 | 6,156.80 | 1,204.92 | 7,361.72 |
| 2001 | 6,415.85 | 1,215.89 | 7,631.74 |
| 2002 | 6,750.43 | 1,472.74 | 8,223.17 |
| 2003 | 6,589.76 | 1,509.03 | 8,098.79 |
| 2004 | 6,406.18 | 1,565.00 | 7,971.18 |
| 2005 | 6,434.99 | 1,805.52 | 8,240.51 |
| 2006 | 6,965.63 | 1,957.87 | 8,923.50 |
| 2007 | 7,630.70 | 2,011.70 | 9,642.40 |
| 2008 | 7,785.02 | 1,928.80 | 9,713.82 |
| 2009 | 7,747.88 | 2,176.29 | 9,924.18 |

说明：联邦政府财政部一个部的信息收集预算占全部预算的80％左右，这是因为，几乎全美每一个个人、家庭都需要报税、退税。财政部收集的信息不是最多，但涉及面却最广，所以给全社会造成的负担也最大。（数据来源：OIRA办公室《年度信息收集预算》）

年年喊减负，负担却年年上升。尽管责任不能全部归咎于政府，OIRA办公室的脸上也挂不住。2010年，OIRA又审时度势、隆重推出了五项减负措施：

1．尽量推行具有自动计算功能的网络填报方式（fillable fileable form）；[03]

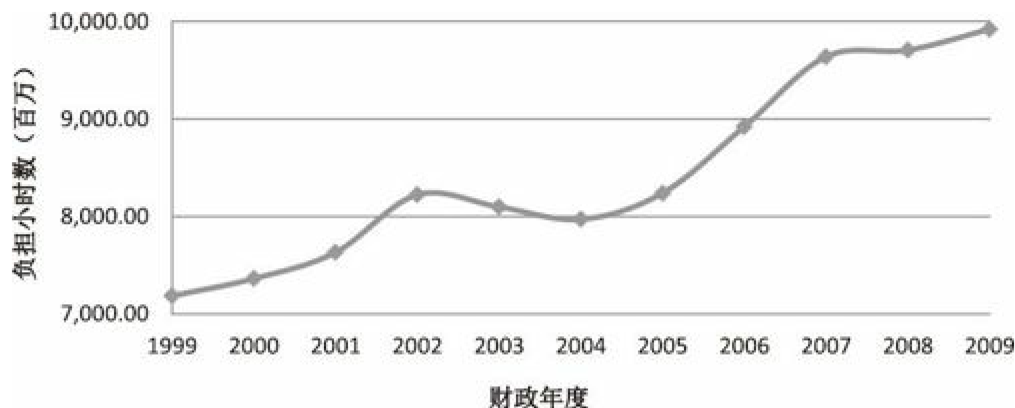
2．在确保安全的情况下，尽量使用电子签名，以减少信息传送、投递的时间；

3．降低信息收集的频度；

4．简化行政过程，避免信息项的重复收集；

5．减少小型企业的信息收集负担。

**1999—2009**年联邦政府信息收集负担的增长趋势




按理说，政府负有管理社会的责任，为完成“公务”，向社会收集信息，听起来天经地义。美国的联邦政府为什么要年年喊减负呢？

这是因为，当今的时代，已经不是“普天之下，莫非王土；率土之滨，莫非王臣”的时代。现代的政府，必须经人民授权产生，它的大政方针，必须符合民意。一个经授权产生的政府，凡事当然都要以授权人的利益为准，凡事都要经得起授权人的质询。

这也是为什么美国政府会把信息收集对社会造成的负担作为衡量信息收集工作重要标准的根本原因。

其实可以预料，OMB关于信息减负的理想难以实现。这是因为，在当前的社会，信息不仅越来越多，其重要性也不断上升，在人、财、物等各个管理要素之间，信息的枢纽作用越来越明显。有效的社会管理，无疑需要更多的信息。

虽然减负的目标不可实现，但OMB关于减负的目标和措施，却是一个经人民选举、授权产生的政府在面对人民时必须具备的姿态。

# 使用法则：隐私，文明社会的共识

隐私，是流淌在美国人血液里的一种价值观。[04]

——安娜·埃舒，美国众议院众议员，2002年

谈到美国的隐私权，我们就绕不开“阳光是最好的防腐剂”这句至理名言的作者：路易斯·布兰代斯大法官。

和奥巴马一样，布兰代斯是哈佛法学院毕业的博士。前文提到，奥巴马在哈佛就读期间，曾经担任过《哈佛法律评论》的主编。这本刊物，在美国法学界的知名度很高，正是布兰代斯1887年在哈佛任教期间创办的。

1890年，布兰代斯与他的同学沃伦（Samuel Warren）就在这本刊物上共同发表了美国隐私权的奠基之作：《隐私权》（Right to Privacy）。在这篇文章中，布兰代斯将“隐私权”定义为“不受别人干扰的权利”。他认为，这项权利是个人自由的起点，只有通过界定这项“人类最广泛、文明人最珍视”的权利，个人的“信仰、思想、情感和感受”才能得到保障。



路易斯·布兰代斯（1856-1941）

布兰代斯是美国历史上第一位担任最高法院大法官的犹太人，也是美国历史上知名度最高的大法官之一。其中的原因，在于他侠肝义胆、热心社会公益事业，在他近30年的律师生涯当中，常常为社会弱势群体代言，被称为“法律界的罗宾汉”、“人民的律师”。（图片来源：维基百科）

这种保障不仅仅意味着个人可以对抗他人对其自由的侵扰，也意味着个人享有不受新闻媒体、政府权力干扰和侵犯的自由。

隐私权的提出虽然仅仅只有100多年，但美国人重视“隐私权”的传统却可以追溯到17世纪的英国。

1647年，英国的法典就规定：

“一个人的房子，对他和他的家人来说，就是他的城堡。”

18世纪中期，英国首相威廉·皮特曾在国会的演讲中说：

“穷人的房子，可能已经破败、摇摇晃晃，风在其中穿梭。但风可以进、雨可以进，英格兰的国王却不能进，他的权力止于这间破房子的门槛。”[05]

美国最早意义上的隐私权，也正是集中在以住宅为代表的物理空间之上的。隐私权意味着一个人可以在自己的城堡中不受监督、不受干涉地发展自己的个性，决定自己的生活方式。

但随着技术的发展，美国社会对于隐私权的保护也在发生深刻的变化，其重心不断转移，经历了从住宅到人、再到信息的转变。

布兰代斯也是最早预见到技术的发展将会不断侵蚀人类隐私的先觉者。

布兰代斯诉讼方法（Brandeis Brief）

布兰代斯的成就很多，他也是美国历史上第一位在法律辩护中大量使用数据的律师。

1908年，布兰代斯为俄勒冈州规定女性每天工作时间不得超过10小时的法案作辩护。他在法庭上出示的辩护书，与众不同、轰动一时：仅仅用了2页的篇幅作法律分析，却用了100多页的篇幅援引各种统计数据，以证明劳动时间过长对女性健康所产生的危害。他的辩

护，最后获得了最高法院9名大法官的一致支持。

这种用数据和事实而不是用法律先例来说明立法必要性和合理性的做法，开创了一种新的法律辩护形式，这种形式被称为“布兰代斯诉讼方法”，其后被法律界迅速采纳。后世的许多重大案例，如争取黑人儿童平等择校权的布朗诉教育委员会案，辩护律师即因为采用了“布兰代斯诉讼方法”，大量引用事实和数据，获得了胜利。

1876年，贝尔发明了电话，这项新的技术极大地方便了人们的交流、推动了社会的发展，但也给隐私权投下了阴影。加上照相机、摄像机等新技术的相继出现，隐私权开始面临很大的挑战：窃听与日俱增，偷拍频频发生，“狗仔队”的跟踪泛滥，警方的秘密监控手段也越来越广泛。布兰代斯等一批先觉者开始意识到，除了住宅，即使在公共场所，个人的隐私权也存在着被侵犯的危险。

1928年，美国发生了隐私权历史上著名的奥姆斯泰德诉美国政府一案。[06]

一位名叫奥姆斯泰德的普通公民涉嫌贩卖私酒，联邦调查局（FBI）的官员在没有获得“搜查证”的情况下通过对其住宅电话、办公电话的搭线监听，掌握了其犯罪证据。奥姆斯泰德因此被定罪。但奥姆斯泰德认为：联邦政府的窃听行为违反了宪法第四修正案对个人隐私权的保护，FBI利用这种手段获得的证据不正当，应当予以撤销，他于是上诉到最高法院。

布兰代斯当时担任美国最高法院的大法官。他支持奥姆斯泰德的上诉。但最高法院的9名大法官最后以5：4的比例驳回了奥姆斯泰德的上诉。5名大法官认为，FBI的秘密窃听没有物理性侵入奥姆斯泰德的住

宅，因此不构成对其隐私权的侵犯。

这场官司之所以著名，是因为布兰代斯大法官作为合审团的少数派，发表了他著名的“异见”（Dissenting Opinion）：

“由于新技术的产生和发展，对隐私权的侵犯已经不需要物理的、强制性的侵入，这种新的侵犯正在以微妙的方式广泛地衍生。这种侵犯即使是国家行为，如果没有合法的审批，也应当被视为违宪。”

奥姆斯泰德虽然最后败诉，但布兰代斯的“异见”却引起了广泛的讨论，对美国社会产生了深远的影响。

1967年，又发生了轰动一时的凯兹诉美国政府一案。[07]这一年，美国最高法院终于以7：1的绝对多数完全采纳了布兰代斯的意见，推翻了对奥姆斯泰德一案的判决。

和奥姆斯泰德一案类似，FBI故伎重施，在公用电话亭搭线窃听了当事人凯兹的谈话，获取了其参与组织赌博活动的关键证据，凯兹随后被定罪。

凯兹以相同的理由上诉到最高法院，最高法院最后宣布：FBI经窃听获得的证据侵犯了公民隐私权，为无效证据，予以撤销。最高法院还在判决中明确：人类的隐私权，不仅仅限于住宅，无论何时何地，即使在公共场所，个人也享有隐私权，对其谈话、通讯的侵犯，就是对其个人隐私领域的侵犯。

凯兹案成了美国隐私权保护从以“住宅”为重心到以“人”为重心的分水岭。

近几十年以来，由于信息社会的兴起，美国社会对隐私权保护的重

心再一次发生了重大的转移。促成这种变化的原因在于政府和商业组织都收集了很多关于个人身份的信息。当个人身份数据（Personal Identity Information）广泛存在于政府、银行、医院、学校、酒店、商场、公司等众多组织当中的时候，每个人的活动其实无时无刻不在被不同组织的数据库“记录”和“监视”，这些数据如果被别有用心地利用和整合，个人的隐私和尊严将不可避免地受到侵害。

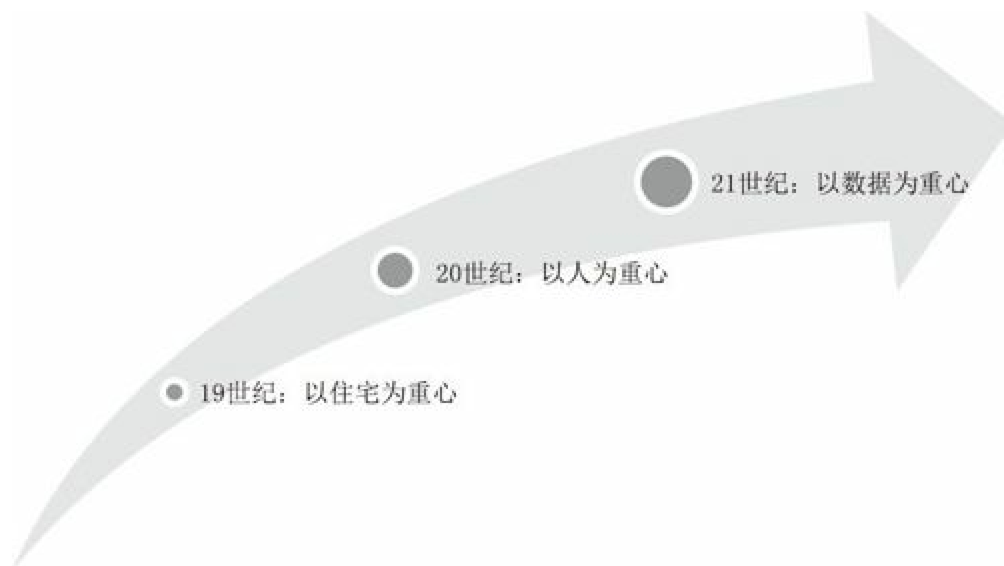
为了应对信息时代的这种冲击，美国对于隐私权的保护，又逐渐从以“人”为重心调整到以“数据”为重心的思路上。

江山代有才人出。这时候，哥伦比亚大学的教授阿伦·韦斯廷（Alan Westin）成了这个新领域的理论先驱和领跑者，他将信息社会的隐私权定义为：

“个人控制、编辑、管理和删除关于他们自己的信息，并决定何时何地、以何种方式公开这种信息的权利”。

1974年，尼克松的水门丑闻全面爆发，行政权力对个人隐私的恶劣入侵，引起了全美朝野上下的反思。此后，美国社会对政府的信任降到了一个历史低点，增加政府的透明度、保护公民的隐私成了全民的共识。在这种情况下，1974年12月，美国国会通过了已经讨论很久的《隐私法》（Privacy Act of 1974）。

美国隐私权保护重心的变迁




韦斯廷教授对数据隐私的研究，为1974年的《隐私法》确定了基本的原则。



阿伦·韦斯廷教授

其著作在西方世界曾经引发跨国性的数据隐私保护运动。代表作有《隐私和自由》（1967年）、《自由社会中的数据银行》（1972年）。

1974年的《隐私法》是一部真正的信息时代的隐私法。它的保护主体就是存储在政府机关内部的“个人信息记录”，如个人的教育经历、工作履历、经济活动、犯罪历史等等记录，它通篇规定的都是美国政府应该如何使用、保护公民的个人信息。

该法规定：行政机关收集保存的公民个人信息，只能用于信息收集时的既定目的；未经本人许可，不得用于其他目的；个人有权知道其信息的使用情况，还可以查询、核对、修改自己被行政机关收集记录的个人信息。

针对如何管理与个人身份隐私相关的数据，美国国会后来还通过了1986年的《电子交流隐私法》（ECPA）、1988年的《计算机查对和隐私保护法》（CMPPA）、2002年的《联邦信息安全管理法》（FISMA）等法律。除了国家层面的立法，美国联邦政府又制定了多个部门规章和实施细则，其中最重要的是1985年行政管理预算局（OMB）颁布并多次修订的《联邦政府信息资源管理政策》，也简称为A-130号通报（OMB Circular A-130）。

该通报明确规定了联邦政府信息管理工作中的“隐私至上原则”：

水门事件（1974年）

在1972年的总统大选中，尼克松竞选团队的5名工作人员闯入位于

华盛顿水门大厦的民主党全国委员会办公室，在安装窃听器并偷拍有关文件时，当场被捕。

尼克松起初推脱说并不知情，并利用总统特权，对调查行为百般阻挠。随着调查的深入，越来越多的白宫官员身陷其中。证据表明，尼克松对此负有直接责任，他长期利用行政资源来收集政治对手的情报。

1974年8月，为了避免弹劾下台，尼克松主动提出辞职，成为美国历史上第一位辞职的总统。

水门事件对1974年《隐私法》和1976年《阳光政府法》的出台都起了直接的推动作用。此后，每当国家领导人发生政治丑闻，便常被新闻界冠之以“门”的名称，如本书后续将介绍的“伊朗门”。

“当联邦政府的信息活动涉及个人信息时，个人隐私权必须切实得到保护。”[08]

“对于个人身份的数据，无论是静态存储，还是动态传输，都必须加密。”

这些信息时代隐私原则的确定，条条款款的背后都有争议，本书将在第七章讲述其中的故事。

当然，并非所有与个人相关的数据都是隐私。对于何种数据才算隐私，要分得一清二楚并不容易，美国的大法官们也没少为这件事纠结，其中最著名的是1972年的联邦政府诉米勒案[09]。通过这场官司，最高法院规定，个人的消费记录不算隐私。

1972年12月，佐治亚州的警察捣毁了一个贩卖、制造私酒的地下工厂，根据掌握的线索，警方怀疑当地的居民米勒是该案的幕后主谋。联邦政府的调查机关给银行发出了传票，要求调阅米勒的银行账户，最后从他账户的交易记录中，掌握了其购买蒸馏设备、运输工具和其他酿酒原材料的一系列证据，米勒因此被定罪。

米勒却拒不认罪，上诉到法院，其律师认为，公民的账户记录属宪法第四修正案保护的隐私范围，必须具有法院下达的搜查令才能获取，政府调查机构的传票没有法定效力，其获得的证据应当撤销。

联邦第五上诉法院支持了米勒的上诉，认为联邦政府违反了程序，侵害了米勒的隐私权。

但联邦政府不服，该案最后上诉到最高法院。

1976年，最高法院宣判，银行的交易记录不属于个人隐私的范围，因为个人的消费记录必须在各个银行、商家之间流动、交换，就像电话号码一样，无法保密，所以不能算是隐私。

米勒最终还是被依法定罪。

# 发布法则：免费，人民已经交税

所有机构因发布信息而收取的费用不能超过其发布信息的成本，不能将收集、处理原始信息的费用包括在内。[10]

——美国联邦政府OMB A-130号通报，1985年

美国政府的信息收集工作以减少社会的负担为目标，其信息使用又以隐私保护为原则，这些规定都和普通大众息息相关，但还有一条几乎和每一个普通公民的经济利益都直接挂钩的法规，你可能没有想到，这条规定，却源于听起来毫不相干的美国《版权法》：

“美国政府的任何工作和作品，都不适用版权保护，但美国政府可以接受并拥有通过转让、赠予或其他方式获得的版权。”[11]

这是1976年美国制定的《版权法》中的第105条。

这短短的一句话一锤定音：对于联邦政府的工作和作品，无论是文字、图像、软件，还是信息和数据，只要是美国联邦政府工作人员为了完成本职工作而取得的成果，都不能申请版权。

当然，美国政府可以接受转让的版权。例如，联邦政府雇用某公司开发一个软件，根据合同，该公司可能拥有该软件的版权，但如果该公司将版权转让给政府，政府则拥有了该软件的版权。此外，美国政府对它所有的工作和作品可以拥有海外的版权。

例如，摄影作品是一种创作，明确受到《版权法》的保护，但只要是美国联邦政府工作人员为完成本职工作而拍摄的照片，一律没有版

权。

有无版权的根本在于收不收费。既然没有版权，政府发布的数据和信息，在法理上就没有收费的理由和根据，只能全民共享、免费发布。[12]

“免费”背后的逻辑其实也很清楚：政府的运行是以公民缴纳的税收为基础的，就好像我们雇请别人来为我们工作，既然已经支付了工资，就不必为其中的某一项服务或工作成果，再付一次钱。

对于如何发布数据，上文提到的A-130号通报（OMB Circular A-130）有具体的规定：任何联邦政府部门收集的数据，必须无偿与其他部门共享，如无法律明禁，还必须向全社会发布。对外发布信息，可以收取信息使用方一定的费用，但这个费用，只能用来弥补由于发布信息这个环节产生的成本，而不能包括数据收集和数据处理等其他环节的成本。

这种收费模式，被称为按边际成本的收费。边际成本，是微观经济学中的一个重要概念。

就政府的信息发布而言，边际成本是指多一个人看到这份信息政府需要投入的成本。例如，如果以CD碟的形式发布信息，多制作一张CD的钱，就是边际成本。如果通过印刷品发布信息，多印一本资料的开支，就是它的边际成本。

互联网产生之后，最方便的信息发布方式当然是政府网站，任何人都可以登陆政府的网站浏览信息和数据，1个人看和100个人看，对政府而言，几乎没有任何区别，那边际成本就将下降到几乎为“零”。

边际成本（Marginal Cost）

边际成本是指在一个特定的产量水平上，增加一个单位产量所需要增加的成本。例如对电脑制造商而言，多制造一台电脑需要投入的人工和原材料。

由于生产的规模效应，随着产量的增加，边际成本可能不断下降。例如，生产第一台电脑的成本可能很高，但生产第100台时，其成本可能就低很多，生产第10000台的时候，其成本可能大幅下降。

边际成本的不断下降正是企业利润的重要来源之一。

和边际成本相对应的经济学概念是“固定成本”。

既然互联网上发布信息的边际成本几乎为零，按照OMB A-130通报的规定，政府就不能收费；对公民而言，就意味着事实上的免费。

2008年，剑桥大学的经济学家波洛克（Rufus Pollock）曾对政府部门发布信息的收费方式做过系统的研究，他总结了3种可能的收费模式。[13]

波洛克指出，欧洲很多国家还在采用第一、第二种方式，政府利用其发布的信息收取费用，以弥补行政经费的不足。这种做法，看起来为政府开源创收了，但由于其对信息的封闭和垄断，对整个社会的经济发展其实弊大于利。而美国政府采用的第三种方式，由于其开放性，将促进全社会的经济增长，尽管这种增长难以量化，但其效果却非常明显。

政府发布信息的3种收费模式

一是利润最大化模式（Profit-maximizing），指的是政府借助自己独家的信息垄断优势，向信息需求方收取市场化的价格。

二是平均成本模式（Average-cost or Cost-recovery）：政府收取一定的费用，这部分费用直接用来支付政府从信息收集到信息发布的所有开支。

三是边际成本或免费模式（Marginal-cost or Zero-cost），指的是只收取信息发布环节所产生的费用，也就是按信息发布的边际成本收取费用。

其中，最为经典的例子是美国海洋和大气管理局（NCAA）发布的天气数据。该局不仅在互联网上为公众免费提供实时的天气预报数据，还提供非常友好的批量数据下载功能。

在世界各国的经济生活中，有不少行业，如农业、交通业、建筑业、保险业、旅游业，无一例外都和天气的变化息息相关。美国商务部曾经做过一个专项分析，最后得出结论，全国有三分之一的GDP产值都和天气情况紧密挂钩。2008年，国家海洋和大气管理局的首席经济学家韦伊尔（Rodney Weiher）估算，仅全国的发电厂，因为得益于该局发布的免费数据，从而可以预测下一天、下一周的发电量、优化资源配置，每年节省的经费达1.66亿美元。[14]

相比之下，英国天气预报数据的管理模式，是公私合营。政府和私营领域的公司合作，政府一方提供原始数据，私营公司则提供数据加工和市场营销，两者通过成立新的实体公司[15]实行利益分成。

英美两国这种不同的模式，曾经引起过广泛的研究和关注，经济学

家普遍认为，虽然难以通过定量研究准确地测量两种模式的效果，但可以肯定的是，免费发布的天气数据是经济发展的巨大助力。2001年，普华永道对此做过专门的调查，它的研究结果是：受益于免费的数据发布模式，2000年美国天气风险管理行业（Weather Risk Management Industry）的产值是整个欧洲的近60倍、整个亚洲的146倍。[16]差距如此巨大，表明了其原因不在“枝节”，而在“根本”。

**2000**年不同地区天气风险行业大小的对比



说明：三个圆形的面积分别表示该地区的天气风险行业大小。

当然，美国这种基于边际成本的免费模式，也并不是每个人都喜欢，事实上，实施起来也有阻力。仅天气数据而言，就有反对的声音。2005年，参议员桑托勒姆（Rick Santorum）提出了《国家天气服务职责法案》（The National Weather Service Duties Act of 2005），该法案主张，除了恶劣天气的预警之外，国家气象部门应该减少免费发布数据的数量，因为这限制了商业竞争。桑托勒姆是2012年共和党党内的总统候选人，他的法案，代表了相当一部分议员的看法，当然，也代表了天气预报产业资本家的利益。但发布信息的边际成本的免费模式已经深入人心，这个法案没赶上全体投票，就在委员会阶段胎死腹中了。

如前文所述，在美国，信息的发布和公开是两个很不一样的概念，发布是政府面向全社会的，公开是点对点、仅仅面对某一特定公民或组织的（具体区别请参见第一章第3节）。但发布信息的边际成本的原则也同样适用于根据《信息自由法》公开的信息和数据。也就是说，当公民根据《信息自由法》提请信息公开时，联邦政府只能收取因为该次查询以及公开这则信息所产生的边际费用，而不能收取政府为了收集、管理这份信息所产生的全部成本和费用。

# 管理法则：质量，互联网时代的根本

数据能满足其既定的用途，它才有质量。如果不能满足既定的目标和用途，就谈不上质量。换句话说，数据的质量不仅取决于它本身，还取决于它的用途。[17]

——杰克·奥尔森，数据库专家，2003年

数据质量的问题，是一个涉及数据收集、使用、发布等所有过程的问题。它的重要性，当然毋庸置疑。

虽然质量的问题贯穿整个数据管理工作的始终，但问题的爆发，常常发生在数据发布的阶段。

自从美国政府成立的第一天起，联邦政府就开始发布数据，但在没有互联网的时代，数据和信息的传播渠道、读者和受众，都相当有限。随着互联网的出现，政府开始在网上发布信息和数据，其受众骤然剧增，传播时间也大大缩短。

这对联邦政府而言，是一个很大的挑战。因为数据一经政府发布，往往被视为权威，对社会的各个领域都可能产生重大的影响。

在互联网时代，任何一份通过网络发布的信息，面对的都不是一个特定群体，而是全体国民。“足够多的眼睛，将使所有的错误都无所遁形。”如果政府所发布数据的质量不可靠，将受到频繁的、大范围的质疑，特别是一些可能会影响到公共政策和行业管制标准的数据，将引起巨大的争议。所以，联邦政府在互联网上发布数据，必须慎之又慎、保

证质量。

为了保证数据发布的质量，OMB曾经制定过3个指导原则：

一是质量要有标准，联邦政府各部门必须制定衡量数据质量的统一标准，这个标准包括3个方面：

客观性：指发布的数据是否准确、客观、可靠；

实用性：是否对社会有用；

完整性：是否能够保证数据在收集、整理过程当中不受到非法的接触和修改。

二是质量管理要有流程。政府各部门必须针对数据质量，完善信息管理的流程，防止低质量的数据出现。

三是要有质量救助机制。政府各部门必须建立一个行政机制来应对社会、公众对于数据质量的质疑和挑战；如果政府发布的数据质量确实存在问题，必须有一个相应的纠错机制来补救。

乍听起来，这3个指导原则都简单、直接，但落实起来，却困难重重。

数据的完整性比较容易界定，但是何为“实用”、“准确、客观、可靠”，就仁者见仁、智者见智，很难有一个统一的标准。在经过几十年的纷争和论证之后，OMB认为：

联邦政府发布的数据，其获得的方式、产生的方法必须是透明的，也就是要向社会公布你的数据收集方法，而且别人通过相同的方法，应

该能够产生、复制相同的数据。

这就要求政府要在发布数据的同时，必须发布一系列的文档，说明数据的来源、产生的方法，以及用户复制过程当中可能出现的问题和错误。

另一个争议的热点在于，谁对数据质量有最终的裁判权？如果所发布数据的质量不过关，联邦政府的纠错机制到底该如何运行？如果政府和社会对某项数据的质量有争议，最终分歧不能统一，是否可以提起法律诉讼？

为了明确地回答这些问题，2001年，美国国会甚至通过了《数据质量法》（Data Quality Act）。但这个法律，还是没有很好地解决问题，它引起了更多的纷争，联邦政府甚至因此多次成为被告，与个人、企业对簿公堂。欲知其中的曲折和详情，请看下章分解。

注释

01 英语原文为：“A popular government without popular information or the means of acquiring it is but a prologue to a farce or a tragedy, or perhaps both. Knowledge will forever govern ignorance: And a people who mean to be their own Governors, must arm themselves with the power which knowledge gives.”—Letter from James Madison to W.T. Barry（August 4, 1822）

02 Information Collection Budget 2010, Office of Information and Regulatory Affairs, P.5.

03 这是美国国家税务局（Internal Revenue Service）率先推出的一种在线报税系统，该系统直接与后台数据库相连，能指导提示用户填入信息，并能自动计算用户需要缴纳的税收，大大减少了用户填报税表的时间。

04 英语原文为：“Privacy is a value that runs through the veins of the American people.”—Anna G. Eshoo, The Honorable Anna G. Eshoo, personal interview, 10 April 2002

05 英语原文为：“The poorest man may in his cottage bid defiance to all the forces of the Crown. It may be frail, its roof may shake; the wind may blow through it; the storm may enter, the rain may enter—but the King of England cannot enter; all his force dares not cross the threshold of the ruined tenement.”—William Pitt

06 Olmstead v. United States, 277 U.S. 438（1928）

07 Katz v. United States, 389 U.S. 347（1967）

08 该法规全文可见于美国白宫网站：http://www.whitehouse.gov/omb/circulars_a130#7，（2011-10-30）。

09 United States v. Miller, 425 U. S. 435（1976）

10 英语原文为："Agencies set user charges for information dissemination products at a level sufficient to recover the cost of dissemination but no higher. They must exclude from calculation of the charges costs associated with original collection and processing of the information."—Circular No. A-130

11 英语原文为：英语原文为："Copyright protection under this title is not available for any work of the United States Government, but the United States Government is not precluded from receiving and holding copyrights transferred to it by assignment, bequest, or otherwise."—Copyright Act of 1976

12 在政府收集的全部信息和数据当中，只有一个例外，这就是1968年通过的《标准参考数据法案》（Standard Reference Data Act）。标准参考数据，指的是某种物质的可以测量的物理和化学属性，这种属性可以用定量的数据表达出来。为物质属性建立标准化的数据指标，是推进科学研究的重要手段。根据该法案，美国商务部可以对"标准参考数据"申请版权，并可以出售，出售的价格可以包括信息收集、编辑、评估、发布以及行政管理的费用和成本。

13 *The Economics of Public Sector Information*. University of Cambridge, Rufus Pollock, November 2008, Available at http://www.rufuspollock.org/economics/papers/economics_of_psi.pdf.

14 *Assessing the Economic & Social Benefits of NOAA Data*, NAS/OECD Conference, Paris. Rodney Weiher. February 2008, at 17-18. Available at http://www.oecd.org/dataoecd/12/31/40066192.pdf.

15 该公司名称为：Weather Exchange Ltd.

16 PricewaterhouseCoopers（2001）The weather risk management industry: survey findings for November 1997 to March 2001. Prepared for the Weather Risk Management Association, June 2001.

17 英语原文为："Data has quality if it satisfies the requirements of its intended use. It lacks quality to the extent that it does not satisfy the requirement. In other words, data quality depends as much on the intended use as it does on the data itself."—*Data quality: the accuracy dimension*, Jack E. Olson, 2003

# 第六章　《数据质量法》的困局

在民主国家，结社的学问是一门主要学问。其余一切学问的进展，都取于这门学问的进展。在规制人类社会的一切法则中，有一条法则似乎是最正确和最明晰的。这便是：要是人类打算文明下去或走向文明，那就要使结社的艺术随着身份平等的扩大而正比地发展和完善。[01]

——托克维尔（1805-1859），法国思想家，1835年

数据的质量问题和隐私安全都曾在美国引起过巨大的风波和争议。

花开两朵，各表一枝。本章先说质量，下章再谈隐私。

上一章说到，1995年，为了进一步规范联邦政府的数据收集管理工作，美国国会修订了1980年通过的《纸面工作精减法》。在这次修订当中，国会要求联邦政府的行政管理预算局（OMB）制定新的政策、拿出具体措施，确保所发布数据的可靠性，即数据要有质量。

但接下来的5年，白宫对此却没有任何的作为和动静，国会最后以失望告终。

在这个问题上，联邦政府为什么迟迟“不给力”呢？

# 产业界“俘虏”政府：数据背后的政经战争

对美国人民来说，最重要的活动就是经济活动。[02]

——卡尔文·柯立芝（1872-1933），第30任美国总统

其实，早在几十年以前，政府所发布数据的质量问题就备受瞩目、广受争议。这是因为，这些数据，往往都关系到行业管制标准的设定。

例如，单位牛奶制品中的蛋白质含量、菌落总数应该是多少？饮用水里能混杂多少含量的微量元素？新鲜蔬菜能带有多少指标的杀虫剂残留？工厂排放的废气、汽车的尾气以及车间的通风条件都要符合相应的标准。

这些标准，都是数据。随着社会的发展、科学的进步，这些标准越来越多、越来越细，每一个都和国民生活和经济发展息息相关。

作为行业管制标准的数据，其大小多少，堪称一个产业的支点，至关重要。每一项新标准的发布、老标准的调整，即使只有零点零几的变化，都可能影响一个行业的竞争、改变一个产业的布局。对于企业来说，这些标准紧一点就可能无利可图，松一点则意味着更大的利润空间。政府的管制，没有最好，就算有，晚一天宣布，企业也可能多赚个十万八千。

天下熙熙，皆为利来；天下攘攘，皆为利往。既然寸“数”寸金，就有人为此大动脑筋。

在华盛顿的国会山上、在宾夕法尼亚西北大道上，每天都穿梭着无

数的说客。他们的任务，就是代表公司、企业和财团游说立法者和执法者，力争把大数变小、小数变无。英语的“说客”一词“lobbyist”，就是指受企业之托、在议会大厅的走道里伺机游说国会议员、焦急等待投票结果的人。

和其他性质的游说活动相比，这种围绕管制标准展开的游说，可谓牵一发而动全身，既敏感又激烈。长期以来，这种游说被美国社会称为政府和企业之间“永不停歇的管制战争”（Ceaseless Regulation War），也是新闻界高度关注的焦点。

这种“战争”，随着政府机构的膨胀、行政权力的扩大，近几十年来进入了“白热化”的阶段。1970年代，美国先后成立了环境保护局（EPA）、职业安全与健康管理局（OSHA）等一批管制型的政府机构，这些机构颁布了大量的行业管制法规。根据联邦日志登记办公室（OFR）的统计，1970年，美国的《联邦管制法规全书》（*Code of Federal Regulations*）总共才54834页，1998年上升到134723页；2009年，又攀升到163333页。这期间，平均每年都要通过近3000项新的管制规定，以2009年为例，该年度通过的管制规定达3503项之多。

虽然联邦政府声称行业管制标准的制定，是以“健全的科学”（Sound Science）为基础的，所引用的数据和研究结果也来源于权威的学术机构和期刊杂志，但产业界对层出不穷的管制规定还是怨声载道。

《联邦管制法规全书》的厚度变化










半个多世纪以来，《联邦管制法规全书》的厚度成倍增长，这既表明了美国政府的管制职能不断加强、一步一步走向了大政府时代，也见证了一场又一场的管制战争。

产业界认为，科学发展无极限，再健全的方法也有瑕疵，再权威的机构也会犯错，政府其实无法确保它们所发布的数据和标准就是建立在“健全的科学”之上的。过多过严的产业管制将提高商品成本、降低工人工资、阻碍技术创新，最终影响经济的发展。更重要的是，政府作为管制者，其管制活动本身也需要被管制，否则就会有滥定标准、滥用数据的可能和危险。

产业界因此要求：为了保证政府的管制权不会被滥用，企业必须能够对行业管制标准背后的数据质量进行质疑和挑战。

正是在1970年代，诺贝尔经济学奖得主乔治·斯蒂格勒（George Joseph Stigler）出版了《管制的经济理论》一书，提出了著名的“政府俘虏理论”（Regulatory Capture Theory）。在考察了联邦政府层面、州政府层面很多个案例之后，斯蒂格勒断言，尽管政府总是号称他们在代表公共利益对产业界进行管制，但现实并非如此，产业界会通过种种手段

来影响政府制定标准的过程，最终俘虏政府、左右监管标准的制定。他的这个理论很好地解释了美国很多管制战争的起因、游说的过程以及最终的结果。



乔治·斯蒂格勒（1911-1991）

1982年诺贝尔经济学奖获得者。斯蒂格勒出生于移民家庭，毕业于芝加哥大学。他主张自由的市场经济制度，反对政府干预，是管制经济学和信息经济学的集大成者，也是芝加哥学派的领袖人物。（图片来源：维基）

面对产业界的抱怨，80年代初，里根政府开始有所表示。里根是著名的共和党人，他倾向于支持产业界的意见。在他执政期间，他签发了一道行政命令，要求所有的机构，如果提出新的管制措施，必须在发布之前通过OMB和他的亲自审查。[03]这开创了总统亲自监控行政管制过程的先河，大大地拉长了行政管制法规出台的周期、限制了政府对商业的

管制。

美国共和党与民主党的主要区别

共和党从1854年建党以来，就一直强调私人企业和个人决策对于促进经济繁荣的重要性，他们长期支持自由的市场经济政策，反对政府管制。正因如此，共和党的候选人一般都会得到美国大财团的支持。

民主党成立于1791年，与共和党相比，民主党更贴近农民、工人和移民，更注重全民福利、劳工权利和民权立法，也因此被称为“平民政党”。民主党强调平等、支持政府加强管制，其候选人多得到工会的支持。

到了1995年，共和党执掌了参众两院，在高涨的企业抱怨和游说态势之下，国会的共和党表示，要从立法的源头上“规治”政府的管制，解决好管制标准背后的数据质量问题。

这时候白宫的掌舵人，是民主党的克林顿。他当然知道，一旦要确保自己所发布数据的质量，无异于给自己的管制权套上一个“紧箍咒”，自缚手脚。此外，数据质量的高低，很难用具体的标准来衡量、判断。于是白宫对国会的要求采取了不置可否的态度。

2000年3月，共和党众议员爱默生（Jo Ann Emerson）又发函催促OMB制定新的措施，保证数据的质量。OMB的主任则再次强调没有必要增加新的措施，回信拒绝了这个要求。

但你有张良计，我有过墙梯。

最终，国会的共和党还是把这个“紧箍咒”戴到了联邦政府的头上。

# 美式“旋转门”：权、名、利大串场

游说在我们体系当中有其正当的位置。我相信支持这个法案的成员都理解游说在美国政治体系中有多么的重要。在不同的时间点、在国会讨论不同问题的时候，每一个美国公民都曾经想要做一名说客。[04]

——比尔·克林顿，第42任美国总统

在签署《1995年游说公开法》时的演讲，1995年

上一章提到，1980年国会通过《纸面工作精减法》之时，在行政管理预算局增设了“信息和管制办公室”（OIRA）。该办公室不仅负责信息的收集工作，还统管行业标准的发布工作。

吉姆·托齐（Jim Tozzi）被任命为OIRA办公室的首任副主任。

在担任该副主任以前，托齐已经在国防部工作了近20年，是一名资深公务员。这次晋升，成了他职业生涯的转折点。因为仅仅3年之后，1983年，他就辞去了这个职务，投身华盛顿的游说业，成了一名在数据质量领域大名鼎鼎的说客。

托齐的说客之路，其实是一名华盛顿说客的标准成长模式。这种模式，在美国被称为“旋转门”（Revolving Door）。

“旋转门”，类似于中国词语里的“走马灯”，指的是一些个人在官场、企业和说客三个角色之间穿梭转换，这边刚刚迈出政府权力机关的大门，那厢就成了资本家的座上宾、大公司的代言人。从权力场到名利场，换汤不换药，转来转去的，都是同一批人。

“旋转门”现象的出现，在美国社会有深刻的背景。进入大政府时代之后，联邦政府的职能不断扩大，对企业的管制不断加强。政府部门需要大量具备专业知识、熟悉企业情况的工作人员来制定行业管制的标准；另一方面，由于政府加大了监管力度，频繁推出更为全面、细致的行业管制措施，企业也迫切需要具有政府部门经验和人脉关系的人员加盟，其中的道理，不言而喻。

这种双向的需要，催生了越来越多的“旋转门”。2005年，非政府组织“公民”（Public Citizen）曾发布一份研究报告[05]，指出从1998年以来有43％的国会议员在离任之后进入了游说产业，成为专职说客，这引起了美国舆论的一片哗然。

很自然，不少学者都对“旋转门”的现象持批评的态度，认为这是国家权力和金钱利益赤裸裸的结合。但问题在于，游说是符合美国宪法的活动，职业转换也是个人的自由。

立国之初，美国就确定了游说的合法地位。此后，在高等法院的若干判例当中，这种权利被不断地明确和加强。例如，最高法院在1853年马歇尔诉巴尔的摩和俄亥俄铁路公司一案的判词中陈述说：

“任何一个人的利益，如果会被一项立法所影响，那他就毫无疑问拥有权利对此进行争辩和游说。他可以亲自辩护，也可以委托专业人士代表他的利益在立法委员会进行游说和在法庭进行辩护。”[06]

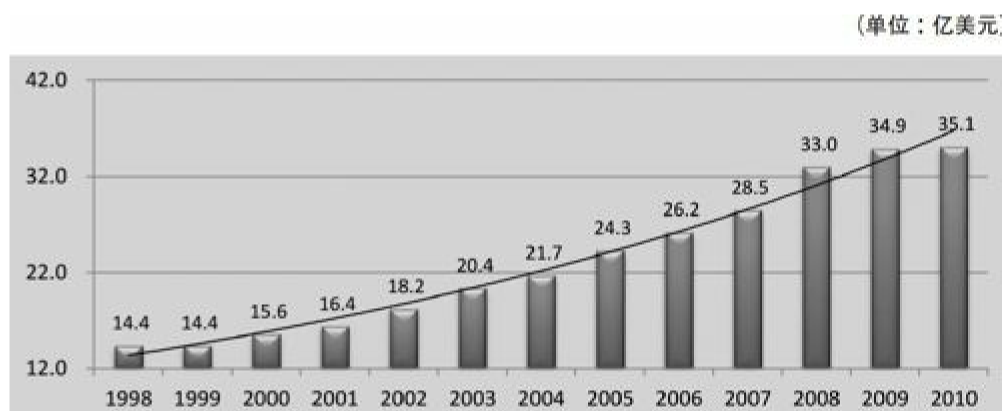
为了减少、控制“旋转门”的现象，美国国会也先后制定了一系列的法律和规定，例如游说人员必须公开注册、政府公职人员必须在一年的冷却期（Cooling-off Period）之后才能进入游说领域。2007年，国会又把离任参议员的冷却期由一年延长到两年。[07]奥巴马对此也极为头痛，

他上任的第一天，除了签署了3个与透明政府相关的法案之外，所签署的另外一个法案就是禁止公务员接受游说团体提供的礼品、禁止高级政府官员在离任2年内从事游说活动。

“旋转门”大行其道，当然是因为游说业利润丰厚、非常景气。

1970年以后，随着游说产业规模的不断扩大，华盛顿的西北角形成了一个“游说产业园”，称为K街（K Street）。在这个街区，遍布着游说公司、公关公司、律师事务所和民间智库。2007年以来，美国的经济一直低迷不振，但K街却依然蓬勃向上、活力四射。根据政治问责中心[08]的统计，2008年华盛顿游说业的规模突破30亿，上升到33亿美金，较2007年上升了13.7％。这个数据还在不断上升。2009年，即使经济持续低迷，花钱最多的前20个游说组织和公司还是增加了他们的游说开支，从2008年的4.18亿上升到5.07亿，多了20％。[09]

**1998**年至**2010**年美国游说业的规模和增长趋势




作为联邦政府的一名高级官员，托齐深谙华盛顿的游戏规则，熟知整个政府管制标准出台的过程及其软肋所在。不难想象，当他决定辞去公职、进军K街之后，他立刻成为炙手可热的人物，被企业界争相聘请。

托齐很快就开始为一家大企业代言，周旋于联邦政府和国会之间，管制标准背后的数据质量问题成了他的“杀手锏”。1995年共和党执掌议会之后，他更如鱼得水，成了提高政府数据质量的主要鼓吹者。2002年，他甚至为共和党的议员起草了《数据质量法》的初稿。

## “掺沙子”法案：国会对付总统的独门秘器

民主，永远不是一件已经完成了的事情。民主是一个过程，需要一个国家永不停懈的努力。[10]

——阿奇博尔德·麦克利什（1892-1982）

美国文学家、国会图书馆馆长，3届普利策奖获得者

第一章提到，对于国会通过的提案，总统有否决权（Pocket Veto）。但根据美国宪法的规定，对于同一个提案，总统不能分项否决，也就是说，他如果赞成其中一部分、反对另一部分，就会面临一个两难选择：要么忍痛割爱，全盘拒绝；要么不管三七二十一，照单全收。

美国国会的议员常常利用这个制度缺陷，在总统表示支持的重大法案里面掺“沙子”，绑定一些令总统头疼的小条款，以逼总统就范。这种“沙子”一般是争议性较大的小问题，独立提交法案，往往难以在总统这边过关。而所谓重要的大法案，常常是总统不得不签的法案，如年度拨款法案，该法案批准联邦政府次年的预算，总统如果不签或者拖延，白宫就有关门的风险。一般情况下，每到年关，总统都会坐如针毡，担心自己的预算案通不过；一旦通过了，即使知道有几粒“沙子”掺杂其中，心下也明白，这就是代价，必须妥协“笑”纳。

这种“沙子”法案，在美国的立法程序中，还有一个更正式的名称，叫“搭便车”法案（Rider）[11]。这个名称也形象地概括了“沙子”法案的另一个特点：临时动议、搭“顺风车”。由于是“搭车”，国会对这种法案一

般都缺乏计划，没有完整的听证过程，也不辩论，往往在临时取得大部分议员的同意之后，就塞进一个大法案里面，匆匆提交总统。

这种做法，一直被批评为美国民主制度的弊病，在历史上曾经多次遭到抨击。1980年代，里根总统在他执政时期曾经强烈要求撤销这种做法。作为受益方，国会当然坚决表示反对。

但1992年，国会的共和党却主动提出，要赋予总统分项否决一个法案中不同条款的权力、取消这种“掺沙子”的做法。

“沙子”法案就好像国会“修理”总统的“独家秘器”，作为三权分立的一极，国会的共和党为什么要主动放权呢？

这话说来也长，又要回到《信息自由法》之争的摩斯时代。

1953年，摩斯第一次提出《信息自由法》的时候，共和党正控制了国会，摩斯寸步难行。到1955年，美国通过中期选举产生了新的第84届国会，那次选举中，共和党完败，民主党成了参众两院的多数党。

没想到共和党从此一蹶不振，此后近40年，一直是众议院的少数党，无法翻身。

这几十年间，共和党也曾多次发起反击，可惜屡战屡败。

长期失去众议院的控制权，共和党的党魁们都深感颜面无存。1992年，为了夺取参众两院的控制权，共和党发起了破釜沉舟的攻势，其全国党部宣布了共和党改革计划（Republican Revolution），抛出了美国历史上著名的“美国之约”（Contract with America），承诺如果成为国会的多数党之后，将推动一系列的改革，取消“掺沙子”的做法就是其中之一。

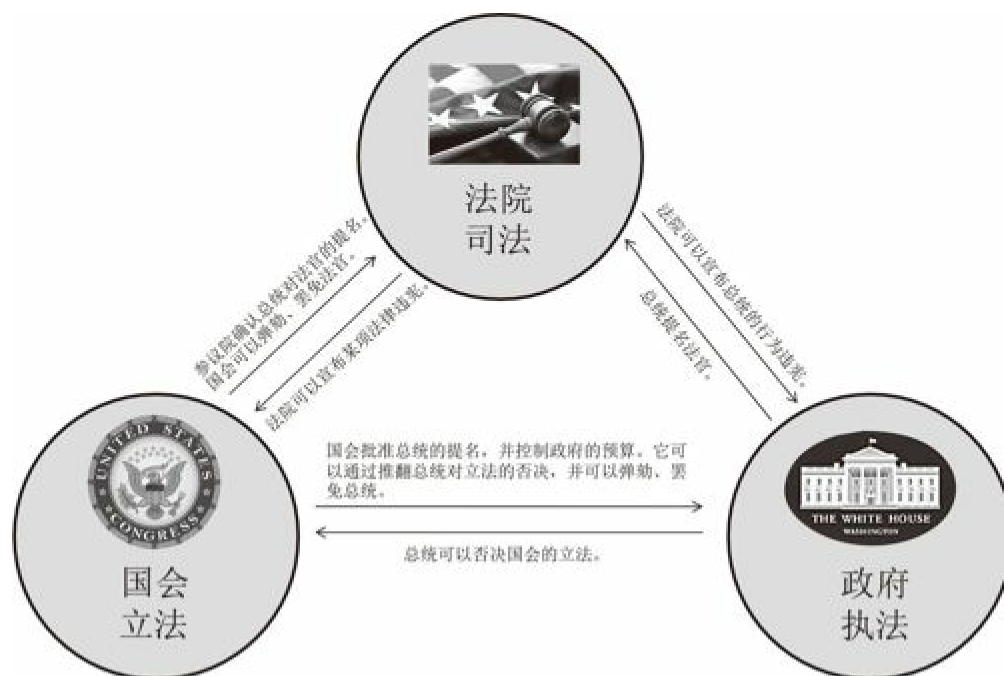
凭借这个大刀阔斧的改革计划，共和党终于如愿以偿，取得了1994年中期选举的全胜，成了参众两院的多数党，40年来首次执掌了众议院。

1996年，他们兑现承诺，推出了《总统分项否决法案》（Line Item Veto Act of 1996），赋予总统剔出“沙子”、分项否决一个法案中不同条款的权力。

时任总统的克林顿当然求之不得，立刻签署了该项法案。

但好景不长。1997年，还是在克林顿任内，这位总统玩大了，他利用这项权力在签发1998年预算案的时候，否定了其中一项关于医院和农民的减税措施，激起了众怒。结果，纽约市表示“伤不起”[12]，一怒之下，该市将克林顿告上了联邦法院，理由是这种做法破坏了三权分立的格局。[13]

三权如何分立和制衡




克林顿虽然是总统，但一审就败诉。他当然不愿意放弃这项权力，随后一直上诉到最高法院。1998年6月25日，最高法院的9名大法官以6比3的多数裁决《总统分项否决法案》违宪，破坏了三权分立的立国之基，必须撤销。

最高法院的判决，就是国家最高、最神圣的法律，国会和政府都不得不从。于是在几年的纷争喧嚣之后，“掺沙子”的做法又静静地回到了原点。

这又引起了另外一个对于美国民主制度的批评和讨论。美国最高法院的大法官是由总统任命的，9位非选举产生的法官是不是应该有权推翻经选举产生的535位国会议员投票通过的国家法律呢？

2006年，小布什主政期间，他也觊觎“分项否决”的权力。在他的倡导下，《总统分项否决法案》经过包装和修改，卷土重来，试图规避最高法院的判决，但也没有在国会冲关成功，最后还是不了了之。

这些问题，可谓是民主制度中的死结，美国争了好几百年，并且反反复复地尝试，都没有定论。这说明，世界上没有任何一项制度是完美的。对于民主制度，英国首相丘吉尔曾经在1947年说了一句大实话：

“我们已经尝试了很多种政府组织的形式，还将在这充满了罪和灾难的世界继续尝试。没有人认为民主制度是完美的或万能的。事实上，民主制度是一种糟糕的制度。但迄今为止，这是我们对政府形式所有的尝试当中发现的最好形式。”[14]

《数据质量法》（Data Quality Act）正是2000年国会共和党在年度拨款法案当中掺的一粒沙子。2000年12月，在共和党众议员爱默生（Jo Ann Emerson）和参议员谢尔比（Richard Shelby）的策划之下，由托齐草拟的《数据质量法》被临时塞进了712页的年度拨款法案（Consolidated Appropriations Act, 2001），成为其第515条附加条款。

这时候的克林顿正急着用钱。在参众两院通过年度拨款法案之后，他看都没有细看，就大笔一挥，欣然签署了这个法案。

所以对OMB来说，《数据质量法》的通过，无异于“暗度陈仓、瞒天过海”，他们头痛不已，却又无可奈何。

因为是粒沙子，该法案总共才两段话、十几行。

“在2001年9月30日前，OMB的主任必须为联邦政府各个部门的数据发布工作制定细则，各部门都必须建立相应的数据审查复核机制，尽最大可能保证联邦政府所发布信息及统计数据的‘质量、客观性、实用性以及完整性’，以落实《纸面工作精减法》。”[15]

虽然才短短百十个字，但这个法案随后却引起了轩然大波。支持也

好，反对也罢，绝大多数人都认为，该法案缺乏足够的酝酿，并不成熟。其中的关键问题，是没有说清楚：谁，对数据的质量拥有最终的解释和裁定权。

这很快得到了证实。

不久以后，这个法案就引发了多宗纠纷以及政府、法院和国会三方之间的分歧。

# 环保“风险门”：公共利益常常无人代表

政府将对工业界进行管制，但这种管制是为了工业界的利益设计和运行的——这是一个规律。[16]

——乔治·斯蒂格勒，美国经济学家，1971年

《数据质量法》颁布之后，联邦政府虽然老大不情愿，但也没有选择。2001年9月，联邦政府各个部门一一制定了部门实施细则，由OMB的信息和管制办公室（OIRA）统一监管各个部门的实施情况。

2002年11月1日，《数据质量法》正式生效。

11月25日，OIRA办公室就收到了关于“数据质量及其客观性、实用性和完整性”的第一宗申诉。

OIRA的前办公室副主任托齐（Jim Tozzi）正是这宗申诉的代理人。

站在托齐背后的，是世界除草剂的行业老大：瑞士的先正达公司（Syngenta）。

先正达公司的主打产品是一种以莠去津（atrazine）为原料的除草剂，近半个世纪以来，它占据了全世界的主要市场，仅在美国的年销售量就达到8000万磅（约36320吨）。

2002年4月，加州大学的海斯博士（Tyrone Hayes）先后在《自然》（*Nature*）、《国家科学院学报》（*National Academy of Science*）等权

威杂志上发表了他对莠去津的研究结果。他认为莠去津在渗入土壤和河流之后，能使雄性青蛙内分泌失调从而转为雌性。这个研究结果，被当时的生物界誉为“突破性的发现”。因为近几十年来，科学家一直搞不清楚，为什么全世界范围内的两栖类动物都在迅速减少。海斯博士的发现有助于解释这个现象。

4月22日，国家环境保护局（EPA）在它的《环境风险评估简报》（Environmental Risk Assessment）上引用了海斯博士的实验结果，认为莠去津存在危害环境的风险。

其实，早在十年前，莠去津的安全性已经开始存在争议。由于它对地表水源的潜在污染和对某些动物的致癌作用，2003年10月，整个欧洲，包括瑞士，都先后禁用了莠去津。但美国的EPA仅仅建议使用者减少用量，并远离水源，这已经为美国的环境保护者所诟病。

对先正达公司来说，海斯的发现可以说是致命一击。不少环境专家都表示，莠去津已经无力回天，必须退出美国市场。

先正达公司当然也一直在为莠去津的安全性进行辩护。颇具戏剧性的是，海斯本人就是先正达公司雇用的科学家。先正达投入巨额资金，支持他的研究，原本希望他得出有利于产品的结果，谁知海斯却在实验中得出了相反的结论，先正达自然悔不堪言。海斯后来还披露，他的研究结果，是在突破了先正达的重重利诱和阻挠之后，才得以正式发表的。

即便如此，先正达还是不甘心放弃自己的市场、坐以待毙。在这个关头，华盛顿的游说业成了他们的救命稻草。这时候的托齐，由于成功游说国会通过《数据质量法》，已经名声在外。

他们一拍即合。

托齐的申诉书说：海斯的实验结果并不可靠，EPA贸然发布其实验结果的做法违反了《数据质量法》。根据该法规定，EPA在刊发这类信息之前，必须检测其引用的实验数据、证明它们是可靠的，并且是可以通过实验重复产生的。EPA并没有履行这个程序，数据质量因此存疑。作为政府部门，其刊发的信息将影响莠去津的正常销售，先正达公司的利益受到了损害。EPA必须对这些数据和结论进行修正。[17]

根据《数据质量法》，EPA随即启动了数据质量的审查复核机制。

经过了近两年冗长的调查，EPA和先正达先后召开了几十场会议。

2004年8月，EPA最终裁决，他们没办法检测和保证海斯博士的数据结果一定“客观、正确”，不能认定莠去津对雄性青蛙的雌化作用。托齐的申诉有效。但EPA同时也要求先正达公司开展更多的产品监测和研究活动。这期间，2004年7月，EPA还发生了轰动一时的“旋转门”事件，主管除草剂项目的副局长费希尔（Linda J. Fisher）辞职，加入了世界第二大化工公司：杜邦公司，成了该公司的副总裁。

先正达转危为安，莠去津继续在美国销售。

这个结果震惊了当时的环境学界。《华盛顿邮报》援引一位环境科学家的话说：这是EPA和先正达之间的一场“妥协和谈判”，而不是基于科学的管制，公共利益受得了极大的侵害。当记者问及“美国那么多的环境保护公益组织，为什么不能让他们代表公共利益参与到审查过程当中来”，EPA主管除草剂项目的主任公开回答说：“那样的话，问题将会变得更加复杂，先正达会拒绝来到会议桌前”。[18]

当然，科学毕竟是科学。2009年11月，海斯博士又发表了更多的研究成果证明莠去津对环境的毒害作用。时隔5年，EPA又重新启动了审查程序，开始了新一轮对莠去津的评估。截至本书成书，审查还在进行当中。

先正达的产品危害的是环境，影响到的人千千万万，但历时十多年，却还是没能扳倒。而且，这不是一宗特例。据《华盛顿邮报》的统计，在《数据质量法》生效的头20个月，OIRA办公室共接受了39宗关于数据质量的申诉，其中有32宗来自公司、行业协会和游说组织。纠纷的重点，都是和公共利益高度相关的行业标准。[19]

在《数据质量法》生效之时，OIRA的主任格雷厄姆（John D. Graham）曾经评论说："无论是公司，还是消费者和环境保护的公益组织，全社会都可以来质疑政府发布的数据是否准确，各种不同的利益，都将得到保护。"但事实证明，这仅仅是冠冕堂皇的说辞；现实是，《数据质量法》仅仅成了商业组织反对管制标准的利器，公共利益的代表常常缺位。

这值得深思。

为什么在美国这样一个高度民主化的国家，在政府与企业游说的博弈当中，公共利益常常还是无人代表或者说缺乏强有力的代表呢？这也恰恰印证了斯蒂格勒指出的："政府最终会被利益集团捕获"。

《数据质量法》的困局，其实反映了人类社会在向前发展过程中遭遇的一个困局。

这种困局，就是个人利益和公共利益之间难以兼顾的矛盾。

# 集体行动的逻辑：人人都想“搭便车”

从政治学的意义上来说，组织起来的少数就是政治上的多数。[20]

——杰斯·杰克逊，美国非裔民权活动家

上文谈到“政府俘虏理论”，斯蒂格勒指出，政府监管部门最终会被产业界俘虏，其监管标准也会被商家所左右。他的这个理论，在20世纪学术研究的历史画卷中，仅仅是重要的一笔，它从属于公共选择理论。

美国的经济学家、社会学家曼瑟尔·奥尔森（Mancur Olson）是公共选择理论的一代宗师。1965年，他出版了《集体行动的逻辑》一书[21]，用经济学的方法剖析了个人与群体的矛盾，他证明了公共利益的代表缺位，是人类社会发展中的困局，每一个社会都难以避免。

集体行动指的是大家一起行动、一致行动，共同承担行动的风险、享受行动的成果。

奥尔森认为，每个人都是经济理性人，也就是说，每个人做事，都会首先考虑自己的利益得失。大家一起行动，一旦行动成功，所有的集体成员都能从中受益，包括那些没有参与行动的、偷懒的、取巧的、旁观的、冷嘲热讽的，甚至使坏的，都将从中获益，而行动的真正参与者、领导者，却可能要付出不同一般的代价。

每个人一计算，都发现自己的最佳策略是“坐等”，是“围观”。也就是说，明知现实不合理，自己也不想出力，都希望别人出头、自己搭车。这种集体都想“搭便车”的结果，就是公共利益得不到有效的照顾、

大家的权益最终都受到损害。奥尔森还指出，当集体越小、利益冲突越激烈的时候，越容易联合；当集体越大、利益冲突并不明显的时候，越难联合。

奥尔森的理论很好地解释了为什么面对同一部《数据质量法》、同一个问题，公司能够迅速地组织起来，进行强有力的游说，而利益受到了侵害的大众，却无法组织起来与公司抗衡。这是因为公司是个小集体，利益集中，容易达成共识，而社会大众的利益高度分散，很难结盟。

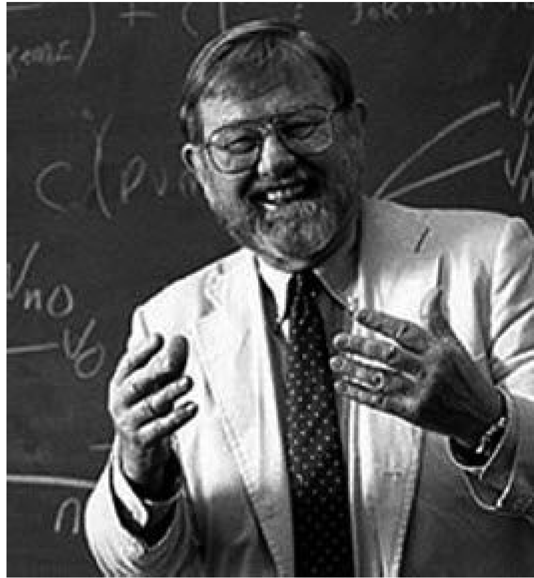
早在2000多年前，亚里士多德就观察到：

“凡是属于最多数人的公共事务，却常常受到最少人的照顾，人们关怀着自己的所有，而忽视公共事务；对于公共的一切，他至多只留心到其中和他个人有些相关的事务。”

面对历史和现实，人类不得不承认：个体是自私的，常常把一己小利摆在公共大利的前面；个体也是有限的，往往只看到一点一线，看不到面，不知道个人与公共的关系，正如点与面的关系。皮之不存，毛将焉附？

千百年来，人的自私和有限，并没有改变。

奥尔森的理论其实也很好地解释了专制制度为什么能在一些国家长期存在。



曼瑟尔·奥尔森（1932-1998）

美国著名的经济学家、社会学家。在他最后一本著作《权力与繁荣》中，他指出："个人权利并不是只有富裕国家才可以负担得起的奢侈品，个人权利是通过以资产和合同为基础的生产获得丰裕回报的根本保证。"[22]（图片来源：马里兰大学IRIS Center网站）

专制者将一个国家大部分人的一部分权利和自由都剥夺了、抹杀了，大家都憎恶专制，但在一定时间、一定程度内，却很少有人反抗。为什么？这是因为，通过反抗，即使能推翻专制制度，其结果是所有人都受益，但出头冒尖的人却可能要付出极大的代价。人的自利天性，使大部分人都选择了沉默，并把希冀的目光投向他人。

专制者当然也懂得这个道理，他们在维护专制的同时，会不遗余力地打击出头冒尖的反抗者，杀鸡儆猴，全力瓦解一切可能发生的集体行动。

但是，奥尔森还是看到了文明的力量。他认为，公共利益的代表缺位、专制社会的困境，最终是可以打破的。

奥尔森的答案是：一个集体、一个社会，要建立合适的激励机制，奖励那些为共同利益作贡献的个人，惩罚那些没有承担集体行动成本的“搭便车者”，从而营造关心公共利益的社会文化和运行机制。

更重要的是，一个社会必须有完善的利益代表机制。美国的游说制度正是典型的利益代表机制，尽管游说加大了政治和金钱相结合的风险，但一定程度的游说，能让公民大众和各种利益团体的正常利益诉求得到充分地表达，这正是民主政治的本质要求。这也是为什么200多年来，美国一直将游说纳入法治的框架之内、不予禁止的根本原因。

要建立一个完善的利益代表机制，有两个关键。

一是授权制度。每个人都有权参与公共事务，但并不是每个人都必须参与公共事务。个人的志趣爱好各不相同，对于公共事务，有人不感兴趣，有人无暇顾及，但他们可以将这份权利授予一个代表，通过代表来行使个人的公共权利。当一个代表接受了一个群体的权利让渡之后，他就可能上升到职业化的高度、提出专业化的意见，他所代表的人群也就成了一个利益相关的集体。这时候，一旦利益受到侵害，他们就很容易联合起来，进行反击。这种机制，一般称为代议制。

二是组团结社的自由。1835年，法国的思想家托克维尔在游历美国之后，出版了轰动一时的《论美国的民主》，被誉为对美国民主制度的经典诠释。他在书中写道：



托克维尔（1805-1859）

法国思想家，26岁时在美国游历，惊讶于美国民主制度的完善，随后将其见闻汇编成《论美国的民主》一书，并在书中探索了美国民主制度的根源。这本书被后世奉为世界政治学史上的经典著作之一。（图片来源：维基百科）

“美国人不论年龄多大，不论处于什么地位，不论志趣是什么，无不时时在组织社团……只要美国的居民有人提出一个打算向世人推广的思想或意见，他就会立即去寻找同道；而一旦找到了同道，他们就要组织社团。社团成立之后，他们就不再是孤立的个人，而是一个远处的人也可以知道和行动将被人们仿效的力量。这个力量能够发表意见，人们

也会倾听它的意见”。[23]

说白了，组团结社就是组成“利益集团”，通过休戚相关的利益，大家紧密相连，放大自己的声音，声张团体的利益，形成一股力量。这股力量，使一个集体可以和其他的组织（包括政府）合作，也可以与它们抗衡。

托克维尔最后概括说：

“在民主国家，结社的学问是一门主要学问。其余一切学问的进展，都取于这门学问的进展。在规制人类社会的一切法则中，有一条法则似乎是最正确和最明晰的。这便是：要是人类打算文明下去或走向文明，那就要使结社的艺术随着身份平等的扩大而正比地发展和完善。”

也就是说，只有通过各种各样的自由结社，各种利益，包括公共利益，才能最终被充分代表。平等和结社，不仅可以帮助人类社会走向文明，也是人类社会保持、延续文明的关键。

# 三权之歧：什么是真正的“和谐”

国家生活的和谐，是各种对抗力量争斗的结果。坦诚表达针锋相对的观点，才会最大可能地以智慧引导政府的行为，打压则常常导致巨大的危险。[24]

——路易斯·布兰代斯，美国最高法院大法官，1920年

《数据质量法》不仅仅面临着公共利益代表缺位的困局。

很快，一宗新的官司，又引发了联邦政府、国会和高等法院之间更多的分歧。

2001年1月，哈佛大学医学院的一位教授在《新英格兰医学杂志》（*New England Journal of Medicine*）发表了他的一个新的发现：减少盐的食用量，可以降血压；即使没有高血压的普通人，如果将每天的食盐摄入量减少到1500毫克，也可以降低患上高血压的风险。

美国卫生研究院下属的国家心肺血研究所（NHLBI）在其网页上宣布了这个发现。

《新英格兰医学杂志》是美国医学界的顶级杂志，心肺血研究所也是国家的权威部门，这个消息发布之后，盐的销量自然应声下跌。

食盐制造行业当然不愿意看到这个结果。

2003年5月，美国盐业所（Salt Institute）和美国商会（CCUSA）两个组织联合向国家心肺血研究所提出申诉，认为“1500毫克”这个数据并

不可靠，要求该所对其进行修正。

该所也启动了数据质量的复审机制。但在几个回合的交涉之后，次年2月，国家心肺血研究所拒绝了修改这个数据的要求。

美国盐业所和美国商会却没有善罢甘休，两个组织随后以《数据质量法》为依据，将国家心肺血研究所一纸告到联邦地方法院，成为《数据质量法》生效之后的首宗法律诉讼。

OMB的官员认为这宗案件开了一个很坏的先例。他们认为，要判断政府发布的数据是否准确可靠，需要大量的专业知识，法院的法官并不具备这个资质。而一些明显违规的企业，将利用漫长的诉讼期作为自己的缓兵之计，拖延有问题甚至有安全隐患的产品停留在市场上的时间。为了避免这种局面的出现，OMB极力主张，政府职能部门应该对“数据质量”有最终的话语权。

但企业界却强调，如果不能诉之于法院，政府发布的数据和行业标准就会成为不可挑战的“金科玉律”，政府的管制权就会成为一种没有监督的权力，就会有被滥用的危险。

国会不少议员也对企业的上诉表示公开的支持，他们表态说，将政府发布数据的质量问题最终置于法律的天平之下，是《数据质量法》立法的初衷和本意。

出人意料的是，2005年11月，弗吉尼亚地方法院却做出了拒绝受理的决定。原告不服，又上诉到联邦巡回法院。2006年3月，第四巡回法院认为：《数据质量法》的相关条款模糊不清，并没有将政府的数据是否“准确可靠”这个问题明确规定在司法的管辖范围之内，原告依据不足，驳回上诉。

针对这个结果，众议院政府改革委员会的主席米勒（Candice S. Miller）评论说，国会立法的本意就是要将政府关于数据质量的决定置于法院的监督之下。她表示这宗案件也许是特例，对其他案件，法院可能会有不同的态度；并表示，如果有必要，国会将考虑修改完善《数据质量法》。[25]

这无异于一种鼓励。

盐业案之后，又有多宗数据纠纷闹到联邦地方法院，也有个别上诉到巡回法院，但无一例外，联邦法院都始终维持了“非司法管辖”的相同观点。尽管如此，个别企业还是前仆后继，再行申诉，新的案件还在审理当中。[26]

政府、国会、法院，是维护美国运行的全部国家机器。但在美国，这三个权力机构常常各行其是、互相指责。这种不团结，听起来贻笑天下，实在令人怀疑它们治理社会的能力，但随着你继续阅读本书，你就会发现，这三种权力之间的争斗，其实是美国的常态。

主张阳光防腐的布兰代斯大法官曾经对这种常态进行过概括，他说：

“国家生活的和谐，是各种对抗力量争斗的结果。坦诚表达针锋相对的观点，才会最大可能地以智慧引导政府的行为，打压则常常导致巨大的危险。”

他认为，恰恰是各种不同力量之间的对抗、各种不同意见之间的争辩，才是美国社会和谐和发展的保证。

注释

01 《论美国的民主》，第640页，商务印书馆，1988，董果良译。
02 英语原文为："The chief business of the American people is business."—Calvin Coolidge
03 该行政命令的序列号：Presidential Executive Order 12291.
04 英语原文为："Lobbying has its rightful place in our system. I believe every Member here and every Member who voted for this bill understands that and understands what a valuable role lobbying can play in the American system. At one time or another, just about every American citizen has wanted to be a lobbyist before the Congress on one issue or another."—President William J. Clinton, Remarks on Signing the Lobbying Disclosure Act of 1995 and an Exchange With Reporters, Dec 19, 1995
05 《从国会到K街之旅》（The Journey from Congress to K Street）。
06 "All persons whose interests may in any way be affected by any public or private act of the legislature have an undoubted right to urge their claims and arguments, either in person or by counsel professing to act for them, before legislative committees as well as in courts of justice."—U.S. Supreme Court，Marshall v. Baltimore & Ohio Railroad Company
07 Honest Leadership and Open Government Act of 2007
08 Center for Responsive Politics是美国一家独立的非政府组织，以收集、公开选举的数据而闻名。
09 这是美国发行量最大的报纸《今日美国》（USA TODAY）的统计结果，Lobbying industry booms in recession，USA TODAY，2/5/2010。
10 英语原文为："Democracy is never a thing done. Democracy is always something that a nation must be doing."—Archibald MacLeish
11 "搭便车"法案（Rider）是指一个追加在其他法案中的条文规定，这个规定和该法案在内容上并没有太大的关联。
12 "伤不起"是2011年前后的网络流行语，意为个体本身早就伤痕累累，已经经不起折腾，经不起伤害。
13 Clinton v. City of New York, 524 U.S. 417（1998）
14 英语原文为："Many forms of Government have been tried, and will be tried in this world of sin and woe. No one pretends that democracy is perfect or all-wise. Indeed, it has been said that democracy is the worst form of government except forms that have been tried from time to time."—Winston Churchill, 1947
15 为行文简洁，文中只翻译了该法的核心部分。该法的部分原文为："The Director of the Office of Management and Budget shall, by not later than September 30, 2001, and with public and Federal agency involvement, issue guidelines under sections 3504（d）（1）and 3516 of title 44, United States Code, that provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies in fulfillment of the purposes and provisions of chapter 35 of title 44, United States Code, commonly referred to as the Paperwork Reduction Act."
16 英语原文为："As a rule, regulation is acquired by the industry and is designed and operated primarily for its benefits．"—The Theory of Economic Regulation, George Stigler, 2 Bell Journal of Economics and Management Science 1,（1971）3-21 at 3
17 REQUEST FOR CORRECTION OF INFORMATION CONTAINED IN THE ATRAZINE ENVIRONMENTAL RISK ASSESSMENT, http://www.thecre.com/pdf/petition-atrazine2B.pdf, Retrieved on 6/28/2011.
18 Data Quality Law is Nemesis of Regulation, Rick Weiss, *Washington Post*, August 16, 2004.
19 32宗申诉案件主要集中在环境保护局（EPA）、野生动物和渔业局（FWS）、国家卫生研究所（NIH）以及消费者产品安全委员会（CPSC）。
20 英语原文为："In politics, an organized minority is a political majority."—Jesse Jackson
21 英语书名为：*The Logic of Collective Action*：*Public Goods and Group Theory*。

22 英语原文为："Rather than being a luxury that only rich countries can afford, individual rights are essential to obtaining the bounteous harvests that property-intensive and contract-intensive production can yield."—*Power and Prosperity*, Mancur Olson

23 《论美国的民主》，第635—639页，商务印书馆，1988，董果良译。

24 英语原文为："Harmony in national life is a resultant of the struggle between contending forces. In frank expression of conflicting opinion lies the greatest promise of wisdom in governmental action; and in suppression lies ordinarily the greatest peril."—Louis Brandeis, Dissent, Gilbert v. Minnesota, 254 U.S. 325, 338（1920）.

25 Unchecked Data: A Tool for Political Corruption? Catherine Campbell Meshkin, Engage Volume 11, Issue 3, December 2010.

26 最新的案件有Americans for safe access v. shalala一案，原告于2010年11月向联邦第九巡回法庭提出重新听证的要求，目前还在审理当中。

# 第七章　全国隐私风波

文明，就是向拥有隐私权的社会不断迈进的进程。野蛮社会的一切都是公共的，靠部落的法则来治理。文明是将一个人从一群人当中解放出来的过程。[01]

——艾茵·兰德（1905-1982），俄裔美国哲学家、文学家，1943年

## 《一九八四》：零隐私的恐惧

不论是睡着还是醒着，在工作还是在吃饭，在室内还是在户外，在澡盆里还是在床上——没有躲避的地方。除了你脑壳里的几个立方厘米以外，没有东西是属于你自己的。[02]

——乔治·奥威尔，《一九八四》第一部第二章，1948年

《一九八四》是英国作家乔治·奥威尔（George Orwell）在1948年创作的一部讽刺小说。

这部小说问世之后，先后被翻译成60多种语言，在世界范围内获得了极高的知名度，被誉为对“极权和专制”进行反思的巅峰巨著。当然，由于其对极权主义入木三分的鞭挞，也在一些国家屡屡被禁止出版。

故事发生在虚拟的大洋国。这里的居民，处于党的完全监视之下，没有任何的隐私。党的领袖就像“老大哥”一样，无处不在、无微不至地“关心”他的国民，确保他们思想“纯洁”、没有任何反叛的行为。一种被称为“电幕”的高科技视频器材被用来监控人民的表达、交流和生活。每个人的住所、办公室和公共场所都安装着监控镜头和窃听器，到处都贴着标语：“老大哥在看着你！”记者们写的一切东西都要经过党的审查才能发表，孩子们也被党煽动、鼓励向警察报告他们父母的可疑行为。

即使一个对党不敬的细微表情，都可能被报告，然后是教育、逮捕、改造、清洗。

故事的主人公史密斯是一名有独立思考能力的青年，他怀疑党的理

论，和一名年轻的女性相爱。他们“出轨”的思想和行为很快就被秘密警察发现，在残酷的监禁、审问、拷打和洗脑之后，这对年轻的恋人最终悲伤无奈地选择了放弃，分别背叛了对方。

《一九八四》刻画了一个无比强大的党和政府、一个令人窒息的恐怖社会。在这里，人性被扼杀，隐私被践踏，思想受到钳制，生活极度贫乏。任何人读完这部小说，都会倒吸一口冷气，对极权主义不寒而栗。1984年，为了纪念奥威尔，这部小说被改编成同名电影，再一次引起了全世界无数心灵的震颤。

在美国的历史上，“老大哥”这种高度集权的党国体制其实并没有出现过。但片中的“老大哥”，通过技术手段实现了无处不在的监控，这一点，给美国人留下了深刻的印象。科学技术不断突飞猛进，政府权力在扩张和蔓延，这些都是对个人隐私的潜在威胁，美国人担心，两者一旦结合，就会有产生“老大哥”的危险。

《一九八四》电影剧照：老大哥在看着你！



说明：乔治·奥威尔的传世之作还有《动物庄园》，该书也是反乌托邦主义的代表作。鉴于他对人类社会的贡献，1984年，一颗新发现的小行星被命名为奥威尔星。

1960年代以后，随着信息技术的推进、计算机的增多，这种危险引起了越来越多人的关注。他们担心，联邦政府会利用信息技术来侵犯个人的隐私。

这种担心并不是空穴来风。

# 大数据就是“老大哥”：中央数据银行之争

我们有很多小的、独立的信息记录系统。这些系统，就单个而言，它们可能无关痛痒，甚至是很有用的、完全合理的。但一旦把它们通过自动化的技术整合连接起来，它们就会渐渐蚕食我们的个人自由。这才是真正的危险。[03]

——美国隐私研究委员会，1977年

1965年，人类的计算模式还仅仅处在第一个阶段——主机时代，这种危险和担心就开始初现端倪。

那个时候，现在白宫的行政管理预算局（OMB）还叫预算局。

预算局提出了一个简单、大胆、在当时堪称革命性的创新计划。

该局建议，联邦政府应该成立一个统一的“数据中心”，把政府部门所有的数据库连接、集中、整合起来，建立一个大型的数据库。预算局相信，这不仅能节约硬件成本，还能提高数据管理、查询和统计的效率；此外，通过部门之间的数据对接和整合，还可以提高数据的准确性和一致性，减少数据的错误。

预算局甚至为这个计划提出了具体的实施方案：人口普查局、劳工统计局、税务局以及社保局等4个数据密集型部门先行一步，首先将数据库连接起来，其他各个部门的数据库逐步纳入，最终的目标是，以公民为单位，为全国每一个人建立一个数据档案，这个档案将包括每一个人教育、医疗、福利、犯罪和纳税等等一切从摇篮到坟墓的数据记录。

预算局将这个大型数据库称为“中央数据银行”。

普林斯顿大学的高等研究院（IAS）是全世界最顶尖的研究机构之一，它的特点，是可以不屈从任何行政的任务和资金的压力，自主开展纯粹的科学研究。时任该研究院主任的是凯森教授（Carl Kaysen），他盛赞这是一个划时代的计划。凯森发表了专门的可行性报告，指出统一管理不仅能节省运营成本、提高数据的准确性和查询的效率，还将更好地保障数据安全。

听起来有百利而无一弊，中央数据银行的计划得到了行政圈、学术界的一致响应。

经过一年的论证，1966年，联邦政府正式向国会提交了“中央数据银行”的方案，请求拨款、开工，开创新的数据管理篇章。

新闻界也开始报道联邦政府的这个创举。

但没想到，新闻界的报道引起了强烈的社会反弹。这种反弹，最后导致了这个计划的流产。

1967年1月，《纽约时报》发表了著名记者、隐私权专家帕卡德（Vance Packard）的文章《不能告诉计算机》，他写道：

“当政府把我们每一个人的信息和日常生活的细节都装进一个中央级的数据银行，我们将受控于坐在电脑机器前面的那个人和他的按钮。这令人不安，这是一种危险。”[04]

美国公民自由联盟（ACLU）是一个成立于1920年、位于纽约的公益组织，它的目标是利用法律的手段维护公民的权利，隐私权正是ACLU关心的重点。对于中央数据银行的计划，ACLU强烈反对，并发

表了一系列的声明和调查。

哈佛大学也对这个计划开展了专门的民调，其调查的结果表明，56％的美国人担心自己的隐私会受到侵害，明确反对这个计划。

一时间，曾经赢得了各方赞誉的数据银行计划在国会的讨论中陷入了泥沼。

此后，美国国会对此召开了一系列的听证会。1968年，众议院隐私委员会发布了一份报告，作出结论说，该计划无法保证公民的隐私不会受到侵害，不予批准。

但这只是一个中断，甚至仅仅只是一个开始。此后的几十年，随着数据库越来越多，类似于“中央数据银行”的计划不断改头换面，在国会发起冲刺。美国各界对其的讨论也从来没有间断过，并形成了泾渭分明的两派。

支持的一派以经济学家和技术专家为主，这个阵营认为，现代经济依赖于大规模的数据整合和交换，统一集成的中央数据库，将提高经济效率，方便大众的生活，是现代社会发展的必然，是社会进步的不二选择。

反对方自然是隐私至上的信奉者。他们认为，在信息时代，无论是个人的日常消费等琐碎小事，还是事关健康、教育的重大决策，都会在各种各样的信息系统当中留下“数据脚印”。这些“数据脚印”，保存在不同的系统中，可能无伤大雅。但如果建立起中央数据银行，通过数据整合和信息加总，就可以再现一个人生活的轨迹和全景，各个系统之间的数据可以彼此印证、互相解释，个人隐私就无所遁形。

隐私学学者认为，这种信息加总和数据整合，无异于一种监控，准确地说，是一种“数据监控”（Dataveillance），其对个人隐私权的侵害，无异于《一九八四》中的电幕。进入大数据时代之后，甚至有专家提出：大数据就是老大哥！（Big Data is Big Brother）

当然，这两派当中，也不乏改变态度、转换阵营之人。

迪博德（John　Diebold），是美国实业界著名的计算机专家。80年代以前，他一直都是“中央数据银行”计划的强力支持者。

1986年，他公开承认：“中央数据银行”的计划如果实行，将会是一个错误。

迪博德是ATM提款机的主要发明人。促使他发生深刻转变的，是推广ATM提款机过程中发生的一件小事。

迪博德的客户在某地区新安装了一部ATM提款机。但接下来几周的数据记录却非常异常：每天的午夜12点到2点之间，有大量的款项被提取。这有悖常情，该银行担心这涉及诈骗等违规操作，于是雇用了侦探对该提款机进行监控和调查。

侦探很快发现，确实有很多顾客午夜来提款。其原因在于，这个提款机，靠近当地一家色情俱乐部。而顾客提取现金的目的，正是不想在信用卡上留下“不体面”的消费记录。

当该银行了解到真相之后，其经理层忍俊不禁。

很快，一篇新闻报道出现在当地的报纸上：“XX银行知道昨晚谁光顾了妓女”。

迪博德分析说：当你在银行存钱、提款的时候，你留下的信息绝不仅仅是一笔银行交易，其实你还告诉了银行，某一时刻你所处的地理位置。这些信息，很可能会成为你其他行为的解释，从而透露你的隐私。例如，这个提款记录，如果和你当天的通讯、消费、旅行等其他数据记录整合起来，你当天的行踪和作为，就不会有太多的秘密可言。

迪博德后来总结说，在信息时代，计算机内的每一个数据、每一片字节，都是构成一个人隐私的血肉。信息加总和数据整合，对隐私的穿透力不仅仅是“1＋1＝2”的，很多时候，是大于2的。

随着时间的推移，数据整合这种“1＋1>2”的效果，在美国社会表现得越来越明显。进入80年代之后，甚至在最高法院的若干判例中也有所体现。

1988年，哥伦比亚广播公司（CBS）的一名新闻记者根据《信息自由法》向司法部下辖的联邦调查局（FBI）提出申请，要求公开一名犯罪嫌疑人麦迪科（Charles Medico）的犯罪记录。个人的罪案记录原本属于隐私，FBI以此理由拒绝了这个要求。但CBS认为，麦迪科已经被确定参与了有组织的犯罪，并贿赂了一名国会议员，他的犯罪记录涉及公共利益，必须公开。CBS后来汇同新闻自由委员会将司法部告上了法院。[05]

一审、二审双方各有输赢，官司最后打到了最高法院。

最高法院认定这是隐私，应该保护。CBS又提出了新的理由，CBS认为，FBI的犯罪记录只是各个执法部门记录的一个加总，麦迪科的这些信息，都曾经在某个特定的时候公开过，不能算作隐私。

1989年3月，最高法院的9名大法官一致否定了CBS和新闻自由委员

会的抗辩，他们在判词中陈述道：

“在一个有组织的社会里，几乎每一则信息都在不同的时候以不同的形式公开过。但是，就个人隐私而言，不同时期零散地公开和一次性完整地公开，即使内容相同，也有本质的区别。”

虽然数据整合这种“1＋1>2”的性质得到了越来越多人的认同，但中央数据银行计划的支持者仍然大有人在。他们相信，中央数据银行的建设，是经济发展的必然。这不是一个“建”与“不建”的问题，而是一个“何时建”、“如何建”的问题。他们更关心技术层面如何实现，他们知道，在通向数据银行的大道上，美国还有一个关键性的问题，没有得到很好地解决。

这就是个人信息的统一标识问题。

如果要将不同的数据库联接整合起来，各个数据库之间必须要有一座桥梁，也就是一个共同的数据项。这个数据项，用数据库的专业术语来说，就是“主键”（Primary Key）。只有通过一个合理的主键，不同的数据才能方便快捷地联接起来。

这个“主键”，就是每个人独一无二的标识。这个标识，在很多国家，就是身份证号码。

而在美国，还从来没有过全国统一的身份证件及号码。因为没有共同的“主键”，在金融、医疗、教育、福利和公共安全等等不同的数据库中，政府的各个部门以及商业组织都各行其是，使用、设计了不同的标识方法。

这些不同的标识方法，是将不同数据库相连起来的一个障碍，也是

影响中央数据银行建设的一个难题。

# 百年纠结：统一身份证

我们的立场是：现在或者可预见的将来，在美国都不应该建立统一的身份证。就严格的定义来说，社会安全号不能算是身份证，它最多只能算一个近似的身份标识。[06]

——联邦政府卫生、教育和福利部部长咨询委员会的报告

《数据记录、计算机和公民权利》，1973年

统一身份证，也是美国近百年来隐私风波中的一个中心话题。

在美国的历史上，联邦政府从来没有发放过全国统一的身份证件。但在现实生活中，有3个证件，不同程度起到了身份证件的作用。

一是驾驶证。从1903年起，美国开始发放驾驶证，该证件被广泛地用作身份证明。但驾驶证是以州为单位发放的，每个州的证件格式、号码位数都不一样；一个人可以拥有多个州的驾驶证，也可以没有驾驶证。所以严格地说，驾驶证虽然具备了身份标识的作用，却还谈不上统一的身份标志，而且很容易伪造。

二是护照，其格式倒是统一的，号码也是唯一的，并由联邦政府发放，但其发放方式，是以自愿申请为基础的，个人不申请、政府就不发放，除了出入境，日常生活中也很少使用。

三是社会安全号。1935年，罗斯福新政期间，他为了建立社会保障系统，提出为每一个有工作的人员，建立一个社会安全账号。每个月，政府将从其工资中扣除一部分金额，并补贴部分金额，共同存入其社会

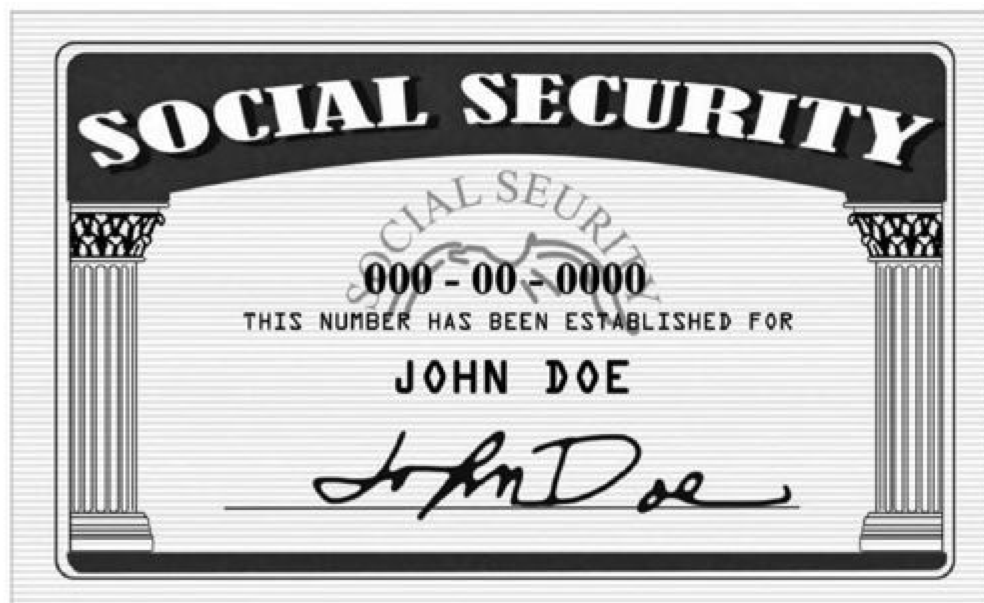
安全账号，以作为其退休养老的保障。

罗斯福这个建立全国统一安全号的提案，也遭到了隐私派激烈的反对。最后，罗斯福向国会妥协并保证：这个号码仅仅用于社会保险，一定不会用于身份标识的领域，并且会被保密。他的这个提案才最终在国会通过。

但一朝天子一朝臣。罗斯福之后，他的承诺并没有被严格地遵守，社会安全号的应用范围还是逐渐扩大了。先是个人纳税的表格启用了这个号码，慢慢地，找工作、申请信用卡、驾驶证、政府福利都需要出示个人的社会安全号。从一定程度上来说，社会安全号似乎也是身份证件的替代品，而且其使用的趋势还在不断上升。

但严格地说，社会安全号还是算不上统一的身份证件。因为它只记录姓名，连男女、年龄、住址、相片等基本的信息也没有，公民也不需要随身携带，该号码被明确规定为个人隐私。1974年的《隐私法》，甚至还有专门的条文规定：“要控制社会安全号对个人隐私造成的威胁”。

美国社会安全卡号的样板




说明：卡上只有个人姓名及社会安全号码，该号码的格式为XXX-XX-XXXX，总共只有9位，前3位、中间2位和最后4位分别代表地区号、群组号和个人序列号。有专家认为，受限于位数，社会安全号不可能成为一个好的身份标识号码。

因为没有统一的身份标识号码，美国各种信息系统中用来标识、区分公民个体的方法自然五花八门，其中最核心的方法，还是人的姓名。但不同的人可以拥有同样的名字，在输入电脑的过程中，还容易出现拼写的错误。毫无疑问，按照名字来查询个体信息的方法，其检索效率和准确率都很低。当要把不同系统的数据库联接整合到一起的时候，要确定“谁”是“谁”，就变得非常困难。

而在具有统一身份证的国家，由于各个数据库之间存在一个共同的、唯一的“主键”——身份证号码，这个挑战就小多了。由于身份证号码在其中穿针引线，各种数据库可以像糖葫芦一样串串相连。

信息技术领域的专家深谙其中的困难。甲骨文公司（Oracle）的创

始人、首席执行官埃里森（Larry　Ellison）就强烈呼吁美国政府建立一套全国统一的身份证系统。他曾经公开表示，如果美国采纳统一的身份证制度，甲骨文公司将免费给联邦政府捐赠一套全国身份证管理的软件。

当然，联邦政府也未尝不想建立全国统一的身份证件。但问题的关键在于：国会不同意！国会的问题又在于：大部分人民不答应！

美国社会的这种争议在第一次世界大战期间就开始了。

1917年，美国对德宣战之后，为了防止德裔美国人在境内从事谍报活动，联邦政府提出，要对德裔这个特殊的群体颁发统一的身份证件、加强管理。还有机构附议：统一的身份证应该延伸到所有的外国人，甚至全体国民。但这遭到了国会的强烈反对。国会仅仅同意了对德裔美国人加强管理，并在最后通过的法案中明确说：身份证的使用只能限于德裔群体，当国际冲突一结束，这种证件就应该立即撤销。

战争一结束，对德裔群体发放的身份证确实就被取消了。但二战期间，美国政府又因为同样的原因，曾经对日裔、意大利裔、华裔群体发放过统一的身份证。[07]

1941年12月，日本发动了著名的珍珠港袭击事件。美国的飞机、舰队受到了日本的重创，共有2402人在袭击中死亡。事件发生不久，盖洛普的民调表明，有69％的美国人支持统一身份证件的做法，明确反对的人降至25％。

但这个数据，只是昙花一现。没过多久，支持者就改变了念头，反对方又在民调中占据了主流。

珍珠港事件之后，联邦政府曾经发起过多次呼吁，要求实施全国统一的身份证件，但都没有在国会取得半数以上的支持，尽数夭折。

1974年，美国司法部成立了一个专门委员会（FACFI），研究统一身份证的问题。该委员会在联邦、州、地方三级政府以及商业领域进行综合调研之后，认为政府有义务为私营领域提供一个统一的标识，以方便全社会各种电子交易行为的发生和管理，因此建议采纳统一身份证的做法。但报告同时坦承：公众对于这个建议的反应是非常负面的。

鉴于此，几乎没有官员敢在公开场合表态支持统一身份证的做法。最后，这个建议只有不了了之。[08]

1967年，“中央数据银行”的争端出现之后，统一身份证更是遭到了“隐私派”的强力反对。

1994年，克林顿执政期间，曾经发起过医疗改革法案，其改革的重点之一，就是为全国每一个人建立一个统一的“医疗服务号”（Universal Health Care Identifier），在这个基础之上，搭建全国统一的医疗数据银行。但这个法案，也被“隐私派”否决了。1996年，克林顿试图重新包装这个想法，让医疗服务号成为另一个法案的一部分（HIPAA），还是没有通过。

近几十年以来，美国也对统一身份证这个问题做过很多次民意调查，但除了珍珠港事件之后的昙花一现，每次都是反对意见占了绝对上风。1942年，民意调查的创始人乔治·盖洛普曾经在《纽约时报》上对统一身份证这个问题做过一个很好的总结。他认为，统一身份证是管理社会、控制公民的一种手段，在美国，这种手段被视为是警察国家的做法，是集权社会的象征，是违反美国精神的。当然，盖洛普也承认，统

一身份证有利于警察打击犯罪、加强国家的安全，特别是在应对恐怖袭击、自然灾害等等突发事件当中，统一身份证将大大方便政府对社会的管理和控制。[09]

统一身份证也多次成为总统大选的辩论话题。前文提到的纽约市的市长朱利安尼，曾经是2008年大选期间知名度最高的候选人，他表态支持全国统一身份证，随后遭到了其他多位候选人的攻击，这些候选人认为：没有任何理由去牺牲公民实实在在的自由以换取“可能会好一点”的国家安全。

## “9·11”大拐点：以反恐的名义向左转

时间在流逝。但美国永远不会忘记“9·11”。我们会记住每一个光荣牺牲的救援队员、每一个生活在悲伤中的家庭，我们会记住这场大火和废墟，记住亲人们的最后一个电话和孩子们的葬礼。[10]

——乔治·布什，第43任美国总统，2001年11月11日

为了国家安全，美国人民必须牺牲一定程度的个人隐私。[11]

——路易斯·弗里，联邦调查局第5任局长，2001年

2001年9月11日，发生了震惊世界的“9·11”恐怖袭击事件。

这天上午，19名来自“基地”组织的恐怖分子先后劫持了4架民航飞机，分别在纽约、华盛顿两个城市实施了自杀式的撞机事件。其中两架撞上了纽约世贸中心的南北双塔，第三架撞向了国防部五角大楼。最后一架据称目标是国会大厦，但由于机组人员和乘客的反抗，飞机最后坠毁在宾夕法尼亚州。

包括恐怖分子，4架飞机上共有256名人员，无一生还。更为悲情的是，时值上班高峰时段，有上万人集聚在世贸中心，两座建筑在遭到撞击之后，先后都着火、燃烧、倒塌；有上千名工作人员来不及疏散，最后都葬身火海，还有近400名消防队员和警察，在大厦里面组织救援，也埋身于废墟之下。

包括劫机者在内，“9·11”总共造成了2996人死亡。这些人，来自全

世界90多个国家。

“9·11”是继1941年珍珠港事件之后，美国遭受的最大袭击，其死亡人数，甚至超出了珍珠港。它对美国人的心理、文化和政治生活都产生了强烈的冲击和影响，在世界历史上，都堪称影响深远的事件。

对“9·11”的调查很快发现，19个劫机者当中，至少有11个人持有虚假的身份证件。

这又重新点燃了统一身份证的大讨论。

曾经全世界最高的建筑成为美国的悲情记忆



世贸中心建于1972年，曾经是全世界最高的建筑，也一直是纽约市最高的摩天大楼和标志。



世贸中心受到袭击轰然倒塌的一刻（图片来源：网络）



说明：在前世贸中心的旧址上，修建了“9·11”纪念碑，碑上刻有全部遇难者的名字。“9·11”十周年的当日，奥巴马夫妇和布什夫妇在新建成的纪念碑前默哀。（图片来源：维基百科）

皮尤研究中心（PRC）随后开展的民调表明，“9·11”事件发生的第二个月，有70％的民众都支持全民统一身份证的计划。

但没想到，这和珍珠港事件之后的社会反应一模一样。仅仅几个月之后，2002年3月，盖洛普的民调表明，统一身份证的支持率又下降到26％，反对率上升为41％。很多人都表示，他们支持在机场加强身份的查验，但反对在银行、医疗、教育、福利等领域使用统一的身份标识。

盖洛普的发言人解释说：

“我们的数据表明，公众只赞成在特定的场合、少数有限的情况下使用统一的身份标识。他们还是怀疑政府会滥用他们的信息。”[12]

但令人意外的是，这一次，统一身份证的讨论并没有持续发酵，中央数据银行的旧计划却再次现身。

9月24日，作为对“9·11”的直接反应，布什政府向国会提交了《爱国者法案》[13]，该法案要求限制公众获取政府信息的广度，并提高政府控制、检查公民个人信息的程度。这些个人信息，很多都属于隐私的范围。例如，根据这个法案，警察和情报机关不需要法院的核准，就有权窃听公民的电话，检查公民的电子邮件和医疗、财务甚至在图书馆的借阅记录等等一切信息记录。



**2001**年**10**月**26**日，布什总统签署《爱国者法案》。（图片来源：**The White House/Eric Draper**）

美国公民自由联盟（ACLU）批评这是向麦卡锡主义的回归，是以“反恐”的名义粗暴侵犯公民的隐私和自由。

但“9·11”的巨大阴影，使整个美国社会的政治文化都开始迅速转向，强调安全、秩序和国家管控的保守主义开始盛行。这时候的美国国会，甚至抛弃了党派之争，在反恐的问题上同仇敌忾、达到了空前的一致。2001年10月26日，《爱国者法案》以前所未有的速度在参众两院高票通过。众议长哈斯特德（John Hastert）解释说，为了打击恐怖主义，确保国家的安全，全体美国人都应该考虑牺牲一部分个人自由。

作为另一个强化国家安全的措施，10月初，布什宣布成立国土安全办公室，统管国家安全事务。很快，布什又向国会建议，将该办公室升格为内阁的部门。

次年11月，国会正式通过了《2002国土安全法》（ Homeland

Security Act of 2002），批准联邦政府在国土安全办公室的基础上成立一个新的内阁部门：国土安全部（DHS）。

2002年11月25日，国土安全部正式挂牌办公，海关、移民、特勤局、应急处理等22个部门、近18万工作人员归入了国土安全部。国土安全部的成立，是1947年以来联邦政府规模最大的一次职能重组，就人数而言，国土安全部成为排在国防部、退伍军人部之后的联邦政府第三大职能部门。

《2002国土安全法》被视为对《爱国者法案》的一个承接。这两份法案，都代表了联邦政府行政权力的迅速扩张和政府透明度的大幅下降，标志着美国联邦政府的历史性转折。

正是在《2002国土安全法》中，中央数据银行计划重出江湖，并再一次引起了全国性的风波和争议。

# 万维信息触角计划：追踪恐怖分子的“数据脚印”

监控就是监控，不管是通过人还是通过计算机来监控，多少都要侵犯个人的隐私和自由。数据库一旦建立，就应该严格保护，否则很容易被滥用。[14]

——西蒙·加芬克尔，美国学者、记者、技术专家，2006年

《2002国土安全法》中重新提出的中央数据银行计划，有一个更响亮的名字：万维信息触角计划（Total Information Awareness）。该法规定，为保证这项计划的实施，首期拨款2亿美元。

为了配合该项计划的实施，2002年5月，国防部在其下属的高级项目研究所（DARPA）内成立一个新的办公室：“信息触角办公室”（IAO）。

这时候国防部的主政者，正是和《信息自由法》几度结缘的强硬派人物：拉姆斯菲尔德。

信息触角办公室新任命的主任，也曾经是全美具有争议的政治人物：原里根时代的国家安全顾问、海军中将约翰·波因德克斯特（John Poindexter）。1990年，波因德克斯特因为身陷“伊朗门”丑闻，被迫辞去公职。这个新的职位，是他沉寂十年之后的东山再起。

伊朗门事件

伊朗门事件是继水门事件之后美国最著名的政治丑闻。

1985 年，联邦政府为了营救被黎巴嫩真主党绑架的人质，违背国会的法律，秘密向伊朗出售军火，以换得伊朗向黎巴嫩真主党的通融。在一名美国人质获得释放之后，白宫高层决定继续交易，波因德克斯特当时担任总统国家安全事务顾问，他同意了该项计划。该计划共换取了3 名美国人质的释放。

1987 年事发之后，波因德克斯特身陷风暴的中心，并涉嫌销毁相关的档案，他在法庭上陈述说“向国会隐瞒真相是其作为政府官员的职责”，引起了全国的哗然。波因德克斯特在司法调查中多次翻供，最后指证该计划获得了里根总统的批准。但里根此时已离任，并患上了老年痴呆症，无法出庭作证。

1991 年，波因德克斯特被无罪释放。

但波因德克斯特万万没有想到，“万维信息触角计划”居然又成了他的“滑铁卢”。短短15个月后，他再度被迫辞职，从此彻底结束了他的政坛生涯。

和1965年相比，2002年，人类的计算模式已经稳步进入了个人型计算阶段，商务智能的各项技术，都已经很成熟。因此，除了将数据联接、集中到一起，提高管理、查询和统计的效率之外，万维信息触角计划还有了新的内容。这个内容，就是数据挖掘。

2002年8月，在国防部高级项目研究所的技术年会上，波因德克斯特首次公开阐述了如何在统一集成的数据库中应用数据挖掘技术，构建“万维信息触角”：

“只有找到新的数据源，反恐工作才能变得更加高效和敏捷。我们

必须把新的和旧的数据库结合起来，从中挖掘出信息，将其转化为知识，并付诸行动。

“这种新的数据资源就是‘交易空间’（Transaction Space）。如果恐怖分子要计划、执行一次恐怖活动，他们必定会在信息空间中留下某种‘数据脚印’。也就是说，他需要‘交易’，这种交易的数据记录，可以是通讯、财务、教育、医疗，也可以是旅行、交通、出入境、房屋等等其他一切数据记录。

我们必须在‘交易空间’中应用数据挖掘的技术，发现和追踪恐怖分子。”[15]




信息触角办公室标志（Logo）

金字塔上一只眼睛射出的光芒，照在地球上代表美国的地方，下面写着“知识就是力量”。这个标志后来备受批评，因为标志中的“眼睛”令人联想到偷窥和监控。

本书第四章曾提到：数据挖掘，是在海量的数据当中通过分析和建

模，发现数据背后隐藏的模式和微妙的关系，以揭示过去的规律、预测未来的趋势。

波因德克斯特大谈数据挖掘，并不是偶然的。从1989年起，美国的计算机协会（Association of Computing Machinery）成立了专门的数据挖掘分会、举办了第一届学术年会、出版了专门期刊，数据挖掘技术就开始在美国受到追捧。此后，政府部门对其的应用也开始逐渐增加。更重要的是，数据挖掘的技术确实不负众望，给很多领域都带来了令人耳目一新的发现。

例如，流感是全球性的传染病。多年来，医务人员一直致力于研究它的爆发周期和特点。1999年，正是通过数据挖掘技术，研究人员获得了突破性的进展。通过对全国2万多个药店的销售数据进行挖掘，科研人员发现，在医院大规模地收治流感病人的两个星期之前，药店柜台的感冒药会有一个销售高峰，这个高峰，只要超过一定的“阀门”值，就预示着一场流感将要爆发。其中的原因在于，人们在患上感冒之后，一般先会尝试自己去买药，直到不见效、症状加重，才会到医院求助，而这个时候，流感已经在社会上全面爆发，失去了最后的治疗控制时机。

哈佛大学的另一位教授通过挖掘病人的就诊数据则发现，由于儿童的抵抗力更弱，儿童的就诊高峰往往是流感的先兆，这个高峰过后的一个月左右，成人流感的爆发就会接踵而来。

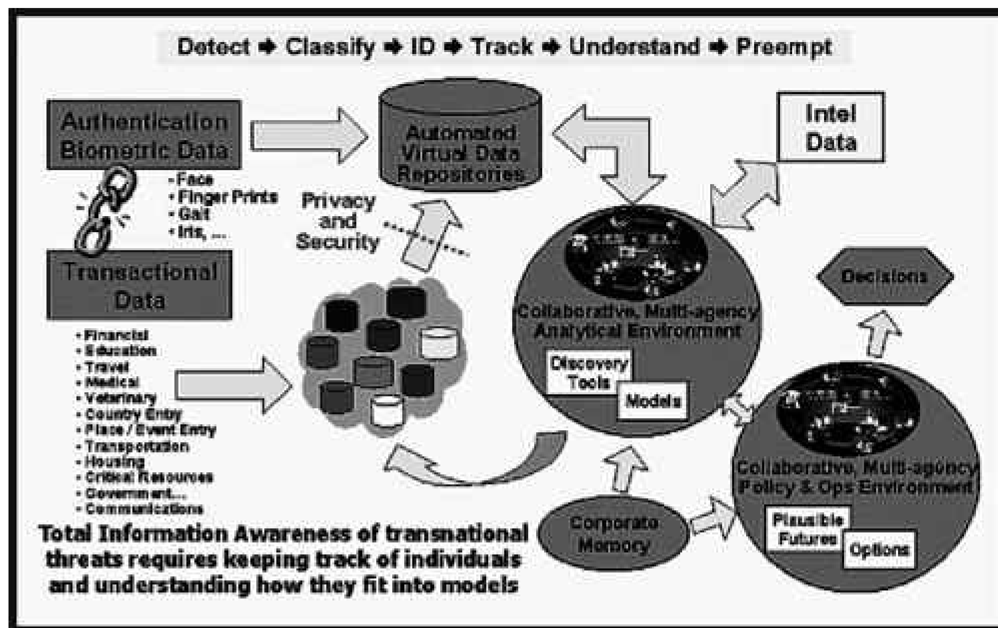
这些发现，都为防治即将爆发的流感，争取到了宝贵的时间。

根据这些发现，匹兹堡大学研发了“疾病爆发实时监控系统”（Real Time Outbreak and Disease Surveillance System）[16]，对宾夕法尼亚州全州药店的药品销售流量进行监控。由于系统的显著效果，2001年11月，

系统的创建人受邀在国会作了专题报告。2002年2月，布什总统亲自到访该实验室，盛赞该监测系统的预警作用。2002年12月，联邦政府疾病预防中心（CDC）开始在全国推广这种数据监控模式。

数据挖掘技术的鼓吹者相信，正像流感的爆发有数据征兆一样，恐怖活动的出现也有迹可寻。因为，和普通人相比，恐怖分子在“交易空间”留下的“数据脚印”有其特定的模式。例如，恐怖分子要经常旅行、流动，他们没有固定的职业和住所，银行的账户上却有充足的现金，他们经常要接听国际电话，要购买特殊的器材、工具，甚至武器。通过在“交易空间”的海量数据中构建一些自动化的数据挖掘“触角”，就可以发现、锁定可疑分子。

波因德克斯特用来介绍“万维信息触角计划”工作原理及流程的幻灯片。




说明：这张幻灯片描绘了个人身份的数据和各种“交易空间”的数据进行整合的过程。（图片来源：维基百科）

“9·11”之后，数据挖掘技术的应用，在美国联邦政府曾经形成一个高潮，同时，也引起了全社会对隐私问题的广泛担心，认为这是一种利用数据对个人活动和交流的变相监控。2003年5月，华裔参议员阿卡卡（Daniel K. Akaka）要求国会问责办公室（Government Accountability Office）对数据挖掘在联邦政府的应用情况进行一个系统的调查。国会问责办公室在对128个联邦部门和机构进行问卷之后，于次年5月向他提交了专门的报告。

报告说，截至2004年5月，联邦政府已经有199个项目正计划使用或已经使用了数据挖掘的技术，其中122个使用了个人信用卡、银行账户、纳税记录等“交易空间”中的个人数据。报告承认，由于数据挖掘技术强大的信息整合能力、分析能力，这已经给公民的隐私保护带来了前所未有的挑战。[17]

万维信息触角计划公开之后，引起了隐私、民权保护团体的强烈反响。美国公民自由联盟（ACLU）指出，所谓的“交易空间”，是关于“所有人、所有事情的记录”，其实无所不包，该组织大声警告：

“如果该系统得以实施，美国人民将生活在《一九八四》所描绘的监控当中，唯一不同的是，监控我们的不是电幕，而是数据库！”

## 6种改变政府的力量：山姆大叔大退让

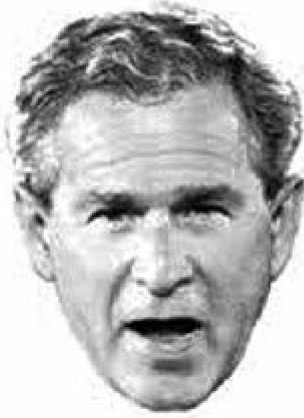
在这个国家，只有存在足够多的人愿意为他们的权利而抗争，我们才能称我们自己是“民主”的。[18]

——罗杰·鲍德温，美国公民自由联盟（ACLU）的创始人

2002年11月24日，著名的作家、新闻记者萨菲尔（William Safire）在《纽约时报》发表文章，批评了万维信息触角计划：

“你用信用卡支付的每一笔购买记录，你订阅每一份杂志，你生病、就诊、吃药的每一份记录，你浏览的每一个网站、发出的每一份邮件，每一笔存款记录，每一次旅行，所有的交易和通讯，都会存储到这个‘虚拟的、中央超级数据库’当中。

这些关于你个人的信息，还仅仅是从商业渠道获得的，再加上政府已经掌握的信息，例如你申请护照、驾驶证提交的表格，你经过收费站的记录，你结婚、离婚的法律文件，你报警的记录等等，这些信息，记录了你的一生。掌握这些信息，曾经是一个超级侦探的终极梦想——这就是现在和我们每一个人都有关的‘万维信息触角’计划。”[19]




美国人为了讽刺布什政府对社会加强监控的做法，制作的布什版“老大哥在看着你！”

一石激起千层浪。

萨菲尔的文章引起了全国的热烈讨论。在他发表文章之前，全国只有12家新闻机构对万维信息触角计划进行了报道。之后，该话题急剧升温，仅仅一个月，各大报纸出现了285篇评论和报道，到2003年2月，全国共出现了508篇相关报道。[20]

除了直接针对万维信息触角计划的分析和报道，各种政府工作人员滥用个人隐私数据的例子也被新闻界挖了出来，作为反面佐证，见诸报端。

密歇根州的一些警察，居然利用警务数据库“猎艳”。他们在街上邂逅漂亮的女性之后，就跟踪她们、记下她们的车牌号码，回到办公室，通过查询数据库，获得该女性的住址等个人信息，然后再人为制造进一步接触的机会。此外，还有警察利用数据库查询政治对手和上司的信息，对相关人员实施威胁。在2001年前后5年期间，涉嫌滥用警务数据库的警察共有90名之多。[21]特拉华州还有政府官员，将公民的个人信息泄露、出卖给赌博公司，帮助他们追债、开发新的顾客。[22]

面对强大的社会舆论，2002年12月，国防部副部长宣布对万维信息触角计划进行审查，确保该项目符合联邦政府的法律和规定。

审查报告很快公布了，国防部在报告中解释说，该项目并没有侵犯公民的任何权利。

但反对的声音仍然持续高涨。先后有3名参议员向国防部写信，质询这个项目。1月10日，国防部再度宣布，将对该项目是否侵犯了公民隐私权进行再次检讨。

2003年2月23日，国会对年度拨款法案进行投票，参议员怀登（Ron Wyden）利用这个机会，在该法案中掺了一粒“沙子”：要求国防部、国家安全局和司法部成立联合工作组对万维信息触角的项目进行调查，并在90天以内向国会提交专门的联合调查报告，否则终止项目的拨款。

作为年度拨款法案中的“沙子”，这个法案当然没有不通过的道理。

5月6日，国会的政府改革委员会又召开了听证会，国防部高级项目研究所的主任在听证会上陈述，该项目的名称已经更改为“反恐信息触角”（Terrorism Information Awareness），专门用于反恐活动，而且，将不使用任何商业机构“交易空间”的数据。

2003年5月19日，在3个部门联合调查报告出台之前，美国公民自由联盟（ACLU）发布了其独立调查报告《对政府超级监控计划的分析》，该报告分为4部分，对国防部前后不一致的解释提出了质疑，指出万维信息触角计划一旦付诸实践，将像“原子弹”一样改变世界。ACLU还指出，“9·11”的发生，并不是政府缺乏情报，而是情报在各个政府部门之间的流动和分享机制不畅通，对情报的分析能力不强造成的。

2003年5月23日，联合调查报告如期抵达国会。在这份报告中，3个部门再次改变口径，强调这是一个“研究性”的项目，充分考虑了公民的隐私权，至于数据，只使用真实的情报信息和仿真数据。

联邦政府步步退让，但万维信息触角计划还是没有逃脱夭折的命运。

2003年8月，波因德克斯特明白大势已去，主动辞去了办公室主任的职务。次月，国会投票通过了新的法案，命令立即终止该项目的开发工作，并停止项目的拨款。该法案还规定，没有国会的授权，今后类似的数据整合挖掘项目，不能使用任何美国公民个人的数据，只能使用外国人的信息和数据。

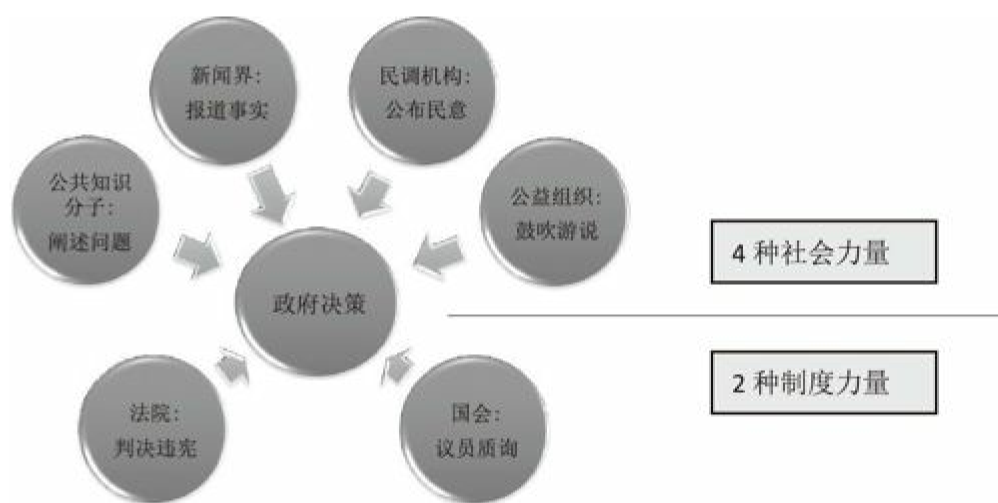
但联邦政府并没有彻底放弃中央数据银行和数据挖掘的计划。

2006年，国土安全部又卷土重来，再以反恐的名义提出了“建言”计划（Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement，简称ADVISE）。这个计划，换汤不换药，还是要在“交易空间”的数据基础上实施数据挖掘技术，以发现潜在的恐怖威胁。这一次，国土安全部声称，他们已经把隐私保护的技术预先嵌入到系统

中，例如，自动屏蔽个人的姓名和社会安全号码。

美国公民自由联盟当然再次披上战袍。这一次，他们在全国发起了一场“只剩6分钟”的动员运动，暗喻离美国变成一个监控性的社会，只有6分钟。

改变美国政府决策的**6**种力量：公共知识分子、新闻界、民调机构、公益组织、国会和法院




《基督科学箴言报》则报道了卡内基梅隆大学数据隐私专家的观点：

“即使没有姓名、没有社会安全号，只要通过性别、生日和邮编3个数据项，数据挖掘的技术就能够成功地识别全美87％的人口。”[23]

2007年9月，“建言”计划再次因为隐私原因，被国会中途叫停。

在中央数据银行计划一次又一次因为美国大众的反对而搁浅的过程中，我们可以清晰地看到，美国联邦政府并不仅仅受限于国会和法院，社会的力量，包括公共知识分子、新闻记者、民意调查机构以及种种公

益组织，也是制约美国政府的强大力量。

2010年9月，美国公民自由联盟发布总结报告说，自“9·11”以来，为捍卫公民隐私，该组织经历了自1920年成立以来最为忙碌的阶段，仅仅为了反击布什政府的《爱国者法案》和《国土安全法》，他们就一共提起了50多宗法律诉讼，有的案件，最后还上诉到最高法院。

据统计，美国公民自由联盟，是除了美国司法部以外在美国最高法院提起诉讼最多的组织。每年，这个公民自治组织有200多名律师和2000多名法律工作的志愿者，处理上千起因为公民维权而产生的法律纠纷。

## 注释

01 英语原文为“：Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled the laws of his tribe. Civilization is the process of setting man free from men.”—Ayn Rand, *The Fountainhead*, 1943

02 本书采用董乐山的译文。

03 英语原文为：“The real danger is the gradual erosion of individual liberties through automation, integration, and interconnection of many small, separate record-keeping systems, each of which alone may seem innocuous, even benevolent, and wholly justifiable.”—U. S. Privacy Study Commission, 1977

04 Don't tell it to the computer, Vance Packard, *New York Time Magazine*, January, 1967

05 U.S. DEPT. of Justice v. Reporters Committee for Freedom of the Press, 489 U.S. 749（1989）, 489 U.S. 749

06 *Records, Computers, and the Rights of Citizens*, Department of Health, Education and Welfare. Secretary's Advisory Committee on Automated Data Systems, 1973

07 二战期间，中国和美国是同盟国，并未有敌对情绪。华裔群体获得身份证是华人主动要求的，因华人担心自己被美国社会当做日本人，主动申请身份证以区别自己。

08 *Card-carrying Americans. Privacy, Security, and the National ID Card Debate*, Joseph Eaton, Rowman & Littlefield Publishers, 1986

09 George Gallup,“Card of Identity for all favored,”*New York Times*, February 4, 1942

10 英语原文为：“Time is passing. Yet, for the United States of America, there will be no forgetting September the 11th. We will remember every rescuer who died in honor. We will remember every family that lives in grief. We will remember the fire and ash, the last phone calls, the funerals of the children.”—President George W. Bush, November 11, 2001

11 英语原文为：“The American people must be willing to give up a degree of personal privacy in exchange for safety and security.”—Louis Freeh

12 Support for ID Cards Waning, Julia Scheeres, Wired News, March 2002

13 所谓的《爱国者法案》，其实是一个代称，该法案全称为“使用适当手段来阻止避免恐怖主义以团结和巩固美国的法案”（Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act），该法案全称的缩写正好是英语中的“爱国者”（PATRIOT），所以被称为“爱国者法案”。

14 英语原文为：“Surveillance is surveillance: whether the surveillance is done by a person or a computer, some kind of violation to personal privacy or liberty has occurred. The database, once created, must be protected with extraordinary measures and is subject to abuse.”—Data Surveillance, Simson Garfinkel, 2006

15 Overview of the Information Awareness Office, prepared remarks for delivery at DARPA Tech 2002, John Poindexter, Aug. 2, 2002

16 关于RODS Laboratory及其项目的介绍，请参见匹兹堡大学的相关网页www.rods.pitt.edu。

17 DATA MINING: Federal Efforts Cover a Wide Range of Uses, Report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget, and International Security, Committee on Governmental Affairs, U.S. Senate, United States General Accounting Office

18 “So long as we have enough people in this country willing to fight for their rights, we'll be called a democracy.”—ACLU Founder Roger Baldwin

19 “You Are a Suspect”, William Safire, *New York Times*, Nov. 14, 2002

20 Safeguarding Privacy in the Fight Against Terrorism, The Report of the Technology and Privacy Advisory Committee, Department of Defense, March 2004

21 Cops Tap Database to Harass, Intimidate: Misuse Among Police Frequent, Say Some, but Punishments Rare, M.L. Elrick, *Detroit Free Press*, July 31, 2001

22 Feds Accuse Newport Manager of Misusing Criminal Database, *Associated Press*, December 6, 2001

23 US plans massive data sweep, *The Christian Science Monitor*, February 9, 2006

# 下　篇　公民故事
# 公民社会与数据互动之双赢时代

争你们个人的自由，便是为国家争自由！争你们自己的人格，便是为国家争人格！自由平等的国家不是一群奴才建造得起来的！

——胡适（1891-1962），中国思想家

惟此独立之精神，自由之思想，历千万祀，与天壤而同久，共三光而永光。

——陈寅恪（1890-1969），中国思想家

# 第八章　数据开放运动

一个人做事情的动机，可以分为三类：一是求生，二是社会生活，三是娱乐。当我们的动机上升到一个更高的阶段时，我们才会取得进步：不是仅仅为了求生，更是为了改变社会，更理想的是——为了兴趣和快乐。[01]

——林纳斯·托瓦兹，Linux系统主要开发者、开源运动领导人，2001年

# 一个新的世界：从软件开源到数据开放

开放的社会是这样一个社会：在这里，人们已经认识到，要多多少少用批判的眼光来看待禁忌，人们会把个人的决定建立在他们自己的才智和掌控之上。[02]

——卡尔·波普尔（1902-1994），英国哲学家，1945年

我们正在进入一个新的世界，在这里，数据可能比软件还要重要。[03]

——蒂姆·奥莱理，奥莱理媒体创始人、技术活动家，2007年

1946年，人类发明了第一台计算机。

计算机一开始就是由软件和硬件共同组成的，但软件和硬件的发展，早期却未能并驾齐驱。相反，在很长一段时间里，软件都是作为硬件的附属而存在的，甚至“软件”（software）这个单词，也远远滞后于“硬件”（hardware）这个单词，直到1958年才正式形成的。

其中的原因，是因为最早的软件对硬件的依附性很强，换一个硬件平台，就可能无法运行；同一个软件产品，即使运行在相同的硬件平台上，其配置不经过修改，也无法销售给其他的客户。

因为可移植性差、通用性低，这时候的软件，大部分都是由硬件制造商一并开发的，作为硬件的一个搭配，免费提供给计算机的购买者。因为免费，软件的原始代码（也称源代码）大部分时候都是公开的。也

正因如此，几乎所有人都相信，软件的开发必须量体裁衣、逐一定制，软件本身难以成为一个大规模销售的独立产品。

但接下来的半个世纪，软件产业却经历了从无到有、跌宕起伏的发展，形成了一个庞大的产业。在这个风起云涌的过程中，一些公司和个人，在其中扮演了极为重要的角色。

1955年，IBM的两名雇员辞职，创办了全世界第一家专门的软件公司：Computer Usage Company（CUC）。尽管走的是“一对一定制”的商业模式，CUC也不断发展壮大，到1967年，已经拥有了700多名雇员、年收入上千万美元，成了上市公司。

在软件业的发展过程中，美国联邦政府也起了举足轻重的作用。由于软件的可通用性低，整个软件产业最早是由个别大项目推动的。这些大项目，基本上都是由联邦政府认购、买单。其中最大的项目之一，是美国国防部和IBM在1959年开始合作的一个防空项目（Semi-Automatic Ground Environment），该项目致力于用计算机软件处理雷达收集的信号，侦察、跟踪天空飞行物的活动。

这个项目，前后跨度近30年，耗资近百亿美元，它不仅帮助IBM成为软件产业的巨头，也为美国培养了一大批程序员。类似的大项目成了程序员的摇篮和“西点军校”。在这些项目中成长起来的技术中坚，后来散布到各个行业，成为各个领域软件开发的带头人。正是联邦政府这些大项目带来的人才优势，几十年以来，美国的软件业一直在全世界占据领先地位。

1969年6月，IBM宣布，将从1970年起停止免费配送软件的做法，软件和硬件将分开定价。这个软、硬件分拆的决定进一步推动了软件行

业的兴起，新的软件公司如雨后春笋般破土而出，用户开始在硬件提供商之外寻找、购买软件解决方案。

1970年代中期，个人电脑问世，随着电脑进入千千万万个普通家庭，软件开始真正具备成为大众化商品的可能性。

但真正把软件产业推进到商业帝国的，却不是IBM，而是一个年轻的后起之秀：比尔·盖茨。

盖茨相信软件行业蕴藏着巨大的商机。1975年，20岁的他从哈佛大学退学，创办了微软公司，致力于软件开发。此后二十年，如滚雪球一般，微软公司迅速扩大，这位辍学者，最后一手造就了一个世界级的软件帝国。

1976年2月，盖茨发表了一封著名的信件：《致爱好者的公开信》。他在信中谴责一些用户没有付费就使用微软公司的软件产品，并断言，这种不尊重版权、非法复制的做法将会阻碍、打击软件业的发展，他质问道：

“有谁会愿意免费、义务地从事如此专业的工作？哪个业余爱好者能投入3年的人力去开发、调试、编写文档，然后免费发布一个软件产品呢？”[04]

盖茨的这封信被后世认为是软件产业的一个重要里程碑。

此后，软件被理所当然地视为一种专利，是私有的，几乎所有的软件公司都开始拒绝公布源代码，商业软件开始大行其道。这种拒绝公开源代码的软件，后来被称为“专有软件”（Proprietary Software）。比尔·盖茨的声明，引领了一场软件商业化的大潮，也导致了专有软件的盛

行。

长江后浪推前浪。在这个迅速崛起、瞬息万变的新生产业里，英雄辈出、各放光彩。这些英雄人物，不仅不拘于学历，更不拘于权威和禁忌，他们倾听内心的声音、服从理想的召唤。

比尔·盖茨固然是一代英雄，但他很快就受到了挑战。他的论断——“没有人能够免费地从事专业性的开发工作”、“软件的开发应该专有、应该封闭”，后来被事实证明是错误的。

新的挑战者具有很强的理想主义色彩。他们反对专有、反对封闭、反对商业化，他们认为封闭性的软件是一种自私、狭隘的做法，束缚了人类的创造性，阻碍了软件技术的传播和交流；他们强调自由、强调开放、强调创新，要求软件开发人员公开原始代码，认为这种公开不仅能提高软件行业的效率和效能，还有利于学习、交流和创造，为人类社会贡献最大的价值。

1983年，麻省理工学院人工智能研究所的一名资深程序员理查·斯托曼（Richard Stallman）发起了著名的GNU项目，目标是通过程序员之间自由的、志愿的协作，开发一套自由的、开放的操作系统，与定价销售的专有操作系统抗衡。1985年，他又成立了自由软件基金会（FSF），为GNU项目提供法律、经济和技术上的支持。这个不可思议的浩大工程，在1991年，因为另一位伟大程序员林纳斯·托瓦兹（Linus Torvalds）的突出贡献，最后成功实现。

自由软件运动和开源运动三剑客



理查·斯托曼：高中时代就开始为IBM编程，发起自由软件运动的时候不到30岁。自由软件运动的诉求，其实不同于后来的开源运动，它强调自由修改而不是免费，但对开源运动产生了巨大影响。



林纳斯·托瓦兹：完成Linux内核第一个版本的时候，还是一名21岁的在校学生。



埃里克·雷蒙：著名的黑客[05]，不到20岁就开始编程，不仅是开源运动的领袖，也是黑客文化的代言人，有多本著作问世。

三个人都是著名的程序员，先后成为自由软件运动、开源运动的领袖。（图片来源：维基百科）

这个免费的操作系统就是现在大名鼎鼎的“Linux”。

从1990年代起，互联网的普及更为轰轰烈烈的软件自由运动推波助澜，这时候的开放原始代码，不仅意味着在互联网上公开自己的代码，还意味着和分散在世界各地的程序员通过网络进行交流和协作。

1997年，埃里克·雷蒙（Eric Raymond）出版了《大教堂与集市》，他在这本书中详细地论述了开放原始代码的好处。他将传统的从上至下

进行控制的软件开发方法比喻为“大教堂”，而通过互联网上公开分享自己的代码，就好像把一件商品放在了集市之上，自由地让别人浏览、评价。雷蒙总结道：

“足够多的眼睛，将使所有的错误都无所遁形。”[06]

雷蒙认为，Linux的核心开发人员林纳斯·托瓦兹就是这种集市模式的首创者，他因此将以上结论称为林纳斯定律（Linus' Law）。

集市模式的关键，不仅在于开放，更在于通过网络的交流、分工和协作。它的提出，不仅对软件开发本身，甚至对互联网时代的商业模式，都产生了深远的影响。例如，维基百科就是按集市模式通过大众自由协作而产生的，而网络版的大英百科全书则是典型的大教堂模式。

随着越来越多的公司和个人采纳了开放源代码的做法，1998年2月3日，主张开放代码和软件自由的各路精英齐聚硅谷，为这场运动正名，最后一致同意将这种发布代码、其他程序员可以对其进行自由修改并重新发布的做法称为“开源”（Open Source）。没多久，一个新的公益组织开源促进会（OSI）诞生了，雷蒙担任了首任主席，该组织很快为开源软件制定了十个标准，这些标准，获得了全世界软件行业的认同。

开源，不仅意味着以开放的姿态进行知识共享，还代表着自由、平等、协作、责任和乐趣等等理念。

林纳斯曾经就开源问题与微软公司的高层展开过激烈的辩论。微软公司批评开源运动严重地破坏了知识产权、扰乱了软件市场，林纳斯则反击微软是一人独大、金钱障目，他后来解释他的初衷说：

“一个人做事情的动机，可以分为三类：一是求生，二是社会生

活，三是娱乐。当我们的动机上升到一个更高的阶段时，我们才会取得进步：不是仅仅为了求生，更是为了改变社会，更理想的是——为了兴趣和快乐。”

这段话，是开源领袖们的价值宣言，也深刻地解释了开源运动为什么能够成功的原因，因为其经典，被称为“林纳斯第二定律”。

在软件开源的运动当中，还有一个人占有历史性的地位，他就是世界闻名的计算机书籍出版商蒂姆·奥莱理（Tim O'Reilly）。奥莱理也是开源运动的坚定支持者，他常常自己掏腰包，赞助开源社区的活动和会议。1999年，他的出版社奥莱理媒体（O'Reilly Media），推出了全世界第一本开源文集：《开源革命之声》[07]。此后，“开源”的概念才开始慢慢走出程序员的群体，成为一个广为人知的概念。



蒂姆·奥莱理

爱尔兰移民，文科毕业生，后来对计算机产生兴趣，成为一名致力于“传播创新知识——改变世界”的计算机图书出版人。他对他旗下出版社的定位是“新技术的转移者”，他也是“Web 2.0”一词的首创之人。（图片来源：维基百科）

伴随着他们的摇旗呐喊，越来越多的开发人员和软件公司开始加入开源运动的行列。进入21世纪以后，开源软件已经深入人心，几乎整个软件业都认识到了开放源代码的重要性及其好处，开源软件的产品也数不胜数、衍伸到各个领域，最早与开源势不两立的微软公司，也开始拥抱开源的做法。时至今日，要想在任何一个行业，找到一个没有开源软件的例子，已经相当困难。

本书第四章曾经谈到，软件是由代码和数据共同组成的，“开源”，指的只是开放代码，并不包括数据。但当开放代码已经成为共识和现实的时候，新一代的创新者，自然又将眼光投向了数据。

考察软件构成的一个角度




虽然同为软件的一部分，但开放数据（Open Data）和开放代码却大不相同。开放代码面向的对象仅仅是程序员，也就是说，它停留在技术的层面；但数据的开放，其涉及面却广得多，它不仅和技术人员相

关，还与数据的来源、性质以及过去和未来的使用人员都息息相关。

开放数据也不仅仅意味着公开数据。要让数据可以重复使用、自由加工，数据公开的格式至关重要。还有，数据不像代码，代码无关个人，数据却可能涉及个人的隐私。

所以，什么样的数据可以开放、应该开放？又以什么形式开放？谁来开放？这些，都是要在开放之前必须回答的问题。

也因为这种种不同，数据开放面临更大的挑战和阻力，数据开放也没有像代码开放一样在商业领域兴起，数据开放的诉求，首先直指公共领域的公共数据，也就是美国联邦政府收集、拥有的数据。

当然，站在这个潮头浪尖的，还是一批年青的技术精英。他们相信，在互联网时代，无论是代码还是数据，只有开放，才能集聚群体的智慧、激发创新的力量、放射出更大的价值，真正推动社会的进步。

2004年，普林斯顿大学的本科生乔舒亚·陶伯拉（Joshua Tauberer）建立了美国民间首个公共数据的开放网站：TrackGov.us。



乔舒亚·陶伯拉

高中时代就开始学习编程，2000年入读普林斯顿大学心理学专业，利用在校的业余时间开发了TrackGov.us。2010年他获宾夕法尼亚大学语言学博士，目前担任公益网站Popvox.Com的首席技术官。

TrackGov.us是一个关于国会立法法案的数据开放网站。前文谈到过美国国会的立法步骤，美国是一个法律大国，在美国国会，任何时候都有一万多个法案处于讨论状态，每一个法案，都涉及千千万万公司和个人的利益。这些法案由谁发起、哪些议员支持、哪些议员反对，投票的结果、修改的次数、通过的时间、法案的内容以及当前的状态等等记录，不仅可以在这个网站上查询，还可以通过RSS邮件跟踪，所有的数据记录，都可以下载、重复使用。

TrackGov.us每月有近百万的点击率。毫无疑问，该网站给学术研究人员带来了极大的方便，也为关心某个法案的公司、个人提供了在线查询跟踪的工具，同时，这些数据还成为新闻界监督政府的利器。

2008年5月，奥巴马、希拉里、麦凯恩3人的选战正酣。根据TrackGov.us的数据，《纽约时报》揭露：从2007年的财政年度始，国会共进行了220次投票，但因为忙于竞选，奥巴马、希拉里各自缺席151次，麦凯恩更甚，缺席164次，高达75％。该报批评说，他们没有履行职责，却照拿工资。[08]

陶伯拉曾经解释自己建立TrackGov.us的初衷：大部分美国人都不了解美国国会的立法程序，他想帮助普通公民增强对于国会的认识，缩小国会运转的真实情况与他们想象之间的差距。

2007年12月，又是在蒂姆·奥莱理的召集下，陶伯拉和其他29名开放公共数据的推动者共聚奥莱理出版社的加州总部。通过两天的会议，他们制定发布了开放公共数据的8条标准和原则。[09]奥莱理在会上感叹说：

“我们正在进入一个新的世界，在这里，数据可能比软件还要重要。”

> 公共数据开放的8大基本原则
>
> 1．数据必须是完整的
>
> 2．数据必须是原始的
>
> 3．数据必须是及时的
>
> 4．数据必须是可读取的
>
> 5．数据必须是机器可处理的
>
> 6．数据的获取必须是无歧视的
>
> 7．数据格式必须是通用非专有的
>
> 8．数据必须是不需要许可证的

这是历史上第一次开放数据的正式集会，也是美国民间第一次尝试建立开放数据的标准和共识。会议将“数据”定义为“一切以电子形式存储的记录”，会议声明：

我们并不决定什么样的公共数据可以开放，我们仅仅为开放制定标准和原则，定义什么才是“开放”。

从这8大原则可以看出，数据开放运动沿袭了软件开源运动的很多诉求，例如通用非专有、机器可处理、获取无歧视等等，数据开放运动可以看做是开源运动在软件领域的一个深化。

但这种深化，已经不仅仅局限在技术领域。开源运动对美国社会乃至全世界的影响，已经渗透到政治、文化、社会生活等各个方面，衍生出开放政治（Open Politics）、开放政府（Open Government）、开放媒体（Open Media）、开放城市等等一系列的运动和口号。

开放，已经成为互联网时代一股不折不扣的浩荡风潮。

这股风潮的触发和形成，其起点还是本书第二章探讨的摩尔定律。正是由于数据存储器的性能不断上升、价格不断下降，人类复制、传播信息的边际成本一降再降，再加上互联网的出现，信息的交流和共享变得前所未有地简单、方便、快捷，人类开放的梦想和渴望才得以一个接一个地实现。

# 总统的雄心：公共财政支出透明

联邦政府掌握和维护的信息是整个国家的资产和财富。[10]

——奥巴马，2009年1月21日

如果我们坐等他人，那改变就不会到来。我们自己——就是我们一直在等的人，我们自己——就是我们所要寻求的改变。[11]

——奥巴马，竞选总统时的演讲，2008年2月5日

正是在陶伯拉发布第一个数据开放网站TrackGov.us的同年，2004年11月，奥巴马赢得了联邦参议员的选举。

次年1月，他踌躇满志，从芝加哥飞往了华盛顿。

有理由相信，初起乍兴的数据开放运动，也进入了他的视野。

因为，作为联邦参议员，奥巴马就任之后成功推出的第一份法案，就是一份有关数据开放的法案。

美国国会参议员的任期，是6年一届。奥巴马在参议员的位置上，只干了3年多。根据TrackGov.us网站上发布的数据，在他短暂的参议员生涯当中，奥巴马作为第一发起人提出过137个法案，但最终成为法律的，却仅仅只有2个。奥巴马还作为合署人和其他议员共同发起过600多个法案，最后通过的也屈指可数。

奥巴马成功交出的第一份答卷，是2006年4月作为主要合署人和共

和党参议员科伯恩（Tom　Coburn）联合推出的《联邦资金责任透明法案》（FFATA）[12]，这个法案后来产生了广泛的影响，也被称为《科伯恩—奥巴马法案》。

这个法案要求联邦政府向全社会开放所有公共财政支出的原始数据，这些数据，包括政府和私营机构的购买合同、公共项目的投资、直接支付以及贷款等等明细。其基本理念和TrackGov.us是一样的：建立一个完整的、专业的公共支出数据开放网站，以统一的格式提供可以下载的数据，以供公众查询使用。

奥巴马是民主党员，科伯恩属共和党，作为一个跨党提案，该法案在参众两院都高票通过，2006年9月，经小布什签署后，成为法律。2007年，根据该法的规定，网站USAspending.gov上线发布，成为美国联邦政府发布公共支出信息的门户网站。

USAspending.gov是个巨大的数据开放网站，可以对联邦政府2000年以来高达3万亿的政府资金使用情况以及30多万个政府合同商所承包的项目进行跟踪、搜索、排序、分析和对比，其数据每两周更新一次。网站上线之后，受到了社会各界的极大好评，获得了“政府搜索引擎”（Google for Government）的美誉。

当然，这个法案的产生和通过，并不仅仅是奥巴马和科伯恩两人力推的结果。



**2006**年**9**月**8**日，布什总统签署《科伯恩**—**奥巴马法案》，左三为奥巴马。（图片来源：网络）

公共支出数据的开放，真正的先行者，还是美国民间的普通公民。

USAspending.gov网站建立的背后，更有一段国会、政府和公益组织3方携手合作的佳话。

上文说到，开放数据的运动兴起之后，首先直指公共数据。有一批社会活动家和公益组织，多次召开民间会议，发起游说活动，督促政府开放公共数据，政府财政支出的数据正是其中的焦点。

“OMB监督”（OMB Watch）就是这其中最执著、强力的行动者之一。这是一个位于华盛顿的公益组织，由加里·巴斯（Gary Bass）于1983年创办。“OMB监督”以监督政府的预算、税收和工作绩效为己任。它的直接监督对象是总统行政办公室下属的行政管理预算局，该局简称为OMB，这也是该组织之所以命名为“OMB监督”的原因。



加里·巴斯

“OMB监督”的创建者，也是信息公开的主要鼓吹者。

“9·11”之后，他仍然坚持政府应该加大信息公开的力度。他在“OMB监督”工作了28年，除了fedspending. org，还在1986年筹建了知情权网络（Right-to-Know Network），该网站为公众提供环境保护和空气质量的数据。1999年以来，巴斯连续10年当选最具影响力的公益领袖（Nonprofit Times Power and Influence Top 50）。

美国的联邦政府其实很早就开始公开发布公共支出的信息，但这种发布是零散的、不系统的，其发布形式可能是新闻发言、文字简报，也可能是电子数据，很不统一。2005年，巴斯意识到，如果能将联邦政府

全部的开支数据统一发布在同一个网站，全社会就可以更好地查询、研究和监督联邦政府的开支和预算。

巴斯后来为这个项目发起了专门的社会募款活动。通过筹款集资、整理数据、开发网站、接洽政府等一系列的工作，2006年，“OMB监督”推出了美国首个公共支出的数据开放网站：Fedspending.org。

Fedspending.org能够逐条跟踪、记录、分析、加总OMB发布的每一笔财政支出。它推出之后，受到了全国舆论的好评，《洛杉矶时报》盛赞其为“财政透明的起点”。

正是受到这个网站的直接启发和帮助，科伯恩、奥巴马发起了《联邦资金责任透明法案》，奥巴马本人，也在这个法案的草拟和推行过程中，和“OMB监督”建立了合作关系。

《联邦资金责任透明法案》通过之后，根据它的规定，行政管理预算局也必须建立一个和Fedspending.org类似的网站，向全社会统一开放联邦政府所有的公共支出数据。

就功能而言，这个待建的政府网站和“OMB监督”已有的网站是相同的。但不同的是，一个是“民营”，一个是“官办”。巴斯在获悉之后，主动走访了行政管理预算局，表明自己愿意提供技术方面的帮助。行政管理预算局的副局长谢伊（Robert Shea）对此大感意外，他是新网站建设的负责人，在他眼里，“OMB监督”就是一只彻头彻尾的“牛虻”，一年到头对他提意见、叮人咬人。对巴斯主动提供的帮助和合作意向，谢伊一不习惯、二不情愿。

这时候，奥巴马从中斡旋，通过他的穿针引线，“官”“民”双方最后形成了合作关系。新生的USAspending.gov网站几乎原封不动地使用了

Fedspending.org的数据库、应用程序接口（API）和在线文档，行政管理预算局因此大幅节约了经费，“OMB监督”也获得了60万美元的报酬，可谓双赢。[13]

这个例子，证明了监督者和被监督者之间的关系也并不总是对立的，两者之间的关系是可以转变的，政府和公益组织之间也可以开展合作。奥巴马对此体会很深，这正是他上任第一天在他首份总统备忘案中阐述的“多方协作”概念：

“政府应该是多方协作的。多方协作让美国人民积极地参与政府工作。各行政部门和机构应利用新的工具、方法和系统，在各部门之间、各级政府之间全面协作。此外，还要与非营利组织、企业和个人进行协作。各行政部门和机构要广泛征求公众的反馈，以评估协作的效果，确定新的协作机会。”

当然，不可否认，奥巴马能站到开放数据的前沿，和他个人的志趣和爱好也有很大的关系。

有评论说，奥巴马是美国历史上对高科技最感兴趣的总统。他上社交网站、用智能手机，并亲自发推特（即微博），拥有1000多万推特粉丝。2008年的总统大选已经成为历史，不少评论家、分析家都认为，正是因为娴熟地利用了信息技术助选，奥巴马才得以入主白宫。

在他的椭圆形办公室里，除了黑人领袖马丁·路德·金的肖像和家人的照片，奥巴马还摆放着不少机械发明的模型。在他书架的顶层，是1877年威廉姆斯（Henry Williams）发明的汽船活叶明轮模型，底层则摆放着1874年皮尔（John A. Peer）设计的齿轮加工机床模型。白宫的文物馆长埃尔曼（William Allman）曾经评论说，这些金属模型反映了奥

巴马的个性和特点：他对“美国的历史、科技以及创新”都有浓厚的兴趣。

《联邦资金责任透明法案》的通过和效果，加上民间的强烈诉求，都激发了奥巴马的雄心。几年的参议员生涯，奥巴马参与联署过数百个法案，涉及社会民生的方方面面，他清楚地知道，公共财政支出的数据只是联邦政府所拥有数据中的很小一部分，可谓九牛一毛、冰山一角。

2007年2月，奥巴马宣布参选总统。11月14日，他作为总统候选人，来到了硅谷山景城（Mountain View），访问了谷歌公司的总部。在和谷歌员工的对话交流中，他再一次谈到了建设USAspending.gov网站的初衷：

“人民知道得越多，政府官员才可能更加负责任。”



奥巴马的椭圆形办公室，其办公桌正后方的桌子上摆满了他与太太及两个女儿的照片，奥巴马曾经在接受采访时解释说，这些照片在提醒

他“为什么要做现在的事情”。



奥巴马的书柜，上下两层各摆着两个机械模型，都是历史上伟大的发明。

（图片来源：www.whitehousemuseum.org ）

他在演讲中说：

“互联网可能是历史上最伟大的开放工具……作为总统，我将把政府的数据用通用的格式推上互联网。我要让公民可以跟踪、查询政府的资金、合同、专门款项和游说人员的信息。为了确保每一个政府机构都能跟上21世纪的标准，我将会任命我们国家的首位首席技术官

（CTO）。”[14]

当他讲到要把政府的数据推上互联网的时候，他的演讲被一阵掌声打断，当他讲到要开创历史、任命“联邦政府首位首席技术官”时，人群中爆发出一阵更为热烈的掌声。

2009年1月21日，奥巴马走马上任，入主白宫。

他果然很快兑现了诺言。

# 数据民主：印裔首席信息官的崛起

这是一场数据民主化的运动，我们正在把信息的力量放到美国人民的手中。[15]

——维伟克·昆德拉，联邦政府首任首席信息官，2010年

2009年3月5日，奥巴马就任不到两个月，就任命了联邦政府的首席信息官；4月18日，他又任命了首位首席技术官。

这是奥巴马创设的两个全新的职位。

奥巴马的创意激起了联邦政府一系列的创新。此后，联邦政府各个部门不仅出现了越来越多的首席信息官，2010年，联邦政府通信委员会（FCC）甚至还率先设置了首席数据官（Chief Data Officer）的职位。

获任联邦首席信息官的是一位年轻的印度裔移民：维伟克·昆德拉（Vivek Kundra）。作为美国联邦政府历史上第一位首席信息官，他立刻引起了全国的关注。

和奥巴马一样，昆德拉的童年几经迁徙。他1974年出生在印度新德里，2岁随父母迁居坦桑尼亚，11岁全家移民美国。他曾经坦言，这些跨国经历加深了他对于不同国家、不同制度的理解。昆德拉后来加入了美国国籍，他说他热爱美国的民主制度，他甚至选择费城宪法中心作为自己举办婚礼的地点。

昆德拉毕业于马里兰大学，获得过信息管理的硕士学位。他毕业之后，经营过几个创业公司，但并不成功。2001年，他决定加入公共领

域。“9·11”事件发生的当天上午，他正在阿灵顿县政府面试“技术主任”的职位。面试被世贸中心的撞机警报打断了，面试官立即打开了电视，第二架飞机正撞向世贸中心的玻璃外墙。屏幕上浓烈的黑烟和火焰令他震惊，也坚定了昆德拉进入公共领域、服务社会的决心。

昆德拉后来回忆道，这特殊的一天，标志着他在公共领域职业生涯的起点。

昆德拉后来还在弗吉尼亚州州政府的技术管理部门工作过。2007年5月，他被华盛顿特区的市长任命为该市的首席技术官。新市长芬蒂（Adrian Fenty）是一位70后，他重视新技术的应用，强调用“透明和开放”来打击腐败。



维伟克·昆德拉（Vivek Kundra）

联邦政府第一任首席信息官，担任此职时不满35岁。




阿尼西·乔普拉（Aneesh Chopra）

联邦政府第一任首席技术官，和昆德拉一样，他也是印度裔移民、70后。

（图片来源：维基百科）

作为首都的首席技术官，昆德拉带领着一支600人的技术队伍、掌握上千万美元的预算。这个位置，成了他真正起飞的平台。

但昆德拉的崛起，并不仅仅是因为这个职位以及其带来的资源。他的机遇，更多的是因为华盛顿的“地利”。

本书第三章曾提到，1994年，纽约街头警察梅普尔利用地图和数据打击犯罪，他设计了“CompStat”的软件。当这个软件横空出世之后，在美国的治安领域获得了巨大的认可，成为一种警务管理模式，迅速被全国各地的警察部门仿效。

随着“CompStat”席卷全国各地的警察部门，有人在动更多的脑筋：能不能把这种数据分析的模式推广到整个城市的管理，而不仅仅局限于治安领域？

奥马雷（Martin O'Malley）就是这个想法的认真尝试者。

他是马里兰州巴尔的摩市的市长。2000年，在他的推动下，该市推出了基于整个城市管理的“CitiStat”。像“CompStat”一样，“CitiStat”一推出，就获得了成功，引起了全国的关注，后来也获得了哈佛大学的美国政府创新奖。

没过几年，奥马雷又当选为马里兰州的州长，他又把这个项目推广到州一级的政府，推出了“StateStat”。

这自然引起了其他一些城市的仿效。

华盛顿特区正是“CitiStat”的追随者。2004年，他们推出了“DCStat”的项目。[16]

为了建设“DCStat”项目，华盛顿特区收集整理了大量的数据。

这些数据才真正成就了昆德拉。他一上任，就给市长芬蒂建议，“DCStat”的数据不仅仅要给城市的管理者分析、使用，还应该对全民开放。在新市长的支持下，昆德拉推出了“城际数据仓库”（Citywide Data Warehouse）项目，把华盛顿特区政府拥有的各类公共数据推上互联网，允许全球任何人访问和下载。昆德拉把这种做法称为“数据民主化”（Democratizing Data）。



奥马雷市长在召开**CitiStat**的管理分析会议，他强调用数据分析的方法推动城市绩效管理。

奥马雷是律师出身，他1999年当选巴尔的摩市的市长之后，第一个政绩就是推行“CitiStat”，他的成功获得了美国很多个城市的关注和仿效。奥马雷连任两届市长，并于2006年当选为马里兰州州长。（图片来源：Center for American Progress）

之所以称之为“数据民主化”，是因为“城际数据仓库”开放的不仅仅是一个专业领域的数据，而是大量与民生相关的数据。2004年陶伯拉建立的TrackGov.us专注于立法，2007年奥巴马推出的USAspending.gov集中在公共支出，昆德拉开放的却是更综合、更广泛的数据，例如学区学位的情况、青少年的犯罪情况、交通事故发生的时间和地段，以及公用场所的租用情况等等与市民日常生活息息相关的数据。而这些数据，曾经仅仅是政府管理人员的内部决策参考，是官员的“专利”，市民大众无权使用。

当然，作为首席技术官，昆德拉的成就并不仅仅局限在城际数据仓库的项目。在他任职的一年半中，他试图在公共管理的各个环节中使用最新的信息技术：他将市政府的微软WORD办公软件升级为可以在线协

同合作的Google在线文档；他要求市政合同的投标、开标过程一律录像，并在Youtube的视频网站上公开。他还组织开发了一个“数字公民广场”，允许居民通过脸谱网站（Facebook）来提交更换驾驶证的申请；下雪天，居民还可以上网实时查看街道积雪的情况、扫雪的进度。

昆德拉的另外一个“地利”在于：他和奥巴马同在国家的政治心脏。作为首都，华盛顿是一个只有60万人口的小城，昆德拉勇于开拓、敢于变革的形象自然引起了奥巴马的侧目。2008年11月，奥巴马取得了总统大选的胜利、拿到了白宫的入场券，他立刻邀请昆德拉加入他的白宫交接团队，为入主白宫做准备。

城际数据仓库项目的主页界面




截至2011年12月，华盛顿特区“城际数据仓库”的项目已经开放了484组数据。（图片来源：data. octo.dc.gov）

昆德拉当然接受了邀请。回到本书的序幕，2009年1月21日，上任第一天，奥巴马签署了他的首份总统备忘案：《透明和开放的政府》。在这份文件中，他阐述了他的执政理念：通过公民参与、多方合作，建设一个前所未有的开放政府。他还在备忘案中命令联邦政府的首席技术官要会同行政管理预算局（OMB）在120天以内制订出一个开放政府的具体行动方案。

120天，也就是从1月21日算起，限期4个月。

接下来的工作，当然是紧锣密鼓。3月5日，奥巴马提名昆德拉担任联邦政府的首席信息官，很快获得了国会的批准。奥巴马在任命书中说：

“首席信息官将和首席技术官密切配合，实现总统关于技术的设想和计划。……我命令昆德拉发扬美国的创新精神，应用科学技术的力量提高政府的绩效、降低政府的开支。作为首席信息官，他将确保我们的政府以最安全、开放和高效的方式运作。”[17]

经过几个月的交接和磨合，昆德拉已经对于总统“关于技术的设想和计划”以及奥巴马的雄心都心领神会。他清楚地知道自己的使命。3月11日，昆德拉上任还不到一周，就召开了联邦政府各部门信息主管的工作会议，拿出了如何提高绩效、增加联邦政府开放性的主体计划。

这就是：为联邦政府建立一个统一的数据开放门户网站——Data.Gov，全面开放政府拥有的公共数据。

美国公共数据开放的编年里程碑



2004
陶伯拉的
TrackGov.us
2006
巴斯的
fedspending.
org
2007
奥巴马的
USAspending.
gov
2008
昆德拉的
城际数据
仓库
2009
联邦政府的
Data.Gov

## Data.Gov：从旗舰初航到保“数”运动

Data.Gov的主要目标是开放联邦政府的数据，通过鼓励新的创意，让数据走出政府、得到更多的创新型运用。Data.Gov致力于政府透明，全力把政府推向一个前所未有的开放高度。它带来的开放将巩固我们国家的民主，提高政府的效率和效能。[18]

——Data.Gov的目标和使命陈述

但昆德拉很快意识到，联邦政府毕竟不是一个小小的华盛顿特区。现在，他要面对的是美国政府15位内阁部长，70多个独立机构的局长、主任和主席。部中有部、局中有局，仅仅一个数据的开放，就可能要经过层层的审批和反复的磋商。

这层层关卡的背后，是一种沿袭了上百年的行政文化。这种文化，虽然看不见、摸不着，却无处不在。

从古至今，所有的政府，都带有保密、封闭的文化基因。虽然一谈到透明和开放，大部分人都会认可这是正确的价值观，但一旦要自己透明、要自己开放，那开放和透明就立刻变成了一种威胁。面对庞大的公共信息，政府首脑的第一反应往往是“安全第一、保密为上”；具体到一线的工作人员，也是“多一事不如少一事”，直接打上“保密”的标签，最为简单省事。

与当年的《信息自由法》、《电子信息自由法》不同的是，这一次并不是国会“逼宫”，而是联邦政府领导层自发的改革。但尽管总统已经表态，联邦政府各个部门的头头脑脑还是对开放数据的做法感到忧虑，

他们表达了各式各样的反对意见，一时甚嚣尘上，争议主要集中在以下几个方面：

1．原始数据之争

是要开放最原始的数据，还是经过加工和解释的数据？如果数据可以加上解释，各部门又应该加上多少程度的说明、解释和观点呢？

昆德拉主张开放原始的数据。但反对者认为：数据是死的，分析是活的；同一组数据，不同的时间、不同的人可能会有不同的解读。如果不加上自己的注解和观点，数据的使用者、分析者可能会得出与政府部门不一样的结论。此外，很多数据属于高度敏感的指标，例如价格消费指数、失业率等等，对于它们的解读，可能会影响一个行业、一个地区甚至整个国家的经济发展走向。开放公共数据的初衷之一是提高公众对于政府的信任，如果出现不同的解读，可能适得其反，影响政府的公信力。那不是搬起石头砸自己的脚吗？

2．数据安全之争

如前文所述，数据整合往往会产生“1＋1>2”的效果。两组分别貌似安全的数据，一旦整合相联，可能会产生意想不到的结果，对国家的安全产生威胁。

反对数据开放的人认为：互联网上的开放，是面对全世界的开放，国际政治错综复杂，国家利益很容易在不知不觉中受到损害。

3．数据质量之争

各个政府部门之间存在职能交叉，收集的数据也不免有所交叉，如果两个部门的数据不一致，应该以谁的数据为准呢？又该如何向公众解

释呢？即使对同一性质、同一类别的数据，新数据还在源源不断地产生和收集，谁来保证新旧数据之间的一致性呢？

有人认为：由于跨部门、跨时间而产生的数据不一致难以消除，如果数据的质量没有保证，数据的开放给政府带来的不是透明和信任，而将是层出不穷的麻烦。

4．数据粒度之争

即使同一组数据，政府也存在多个版本和形式，应该开放哪一组呢？是粗线条的？还是粒度最小、最细的数据格式？

有人主张数据开放宜粗不宜细。

昆德拉主张：构建语言的单位不是句子，而是字母。人们用字母来创建单词、句子和文章。数据也一样，要用最小的粒度把数据呈现给用户，让不同的用户各取所需，无论是警察还是社区居民，自己去决定怎样组合它们。可能的组合是无穷无尽的。这样数据才能发挥全部的潜在价值。

5．数据价值之争

必须首先开放社会最需要的数据，即高价值的数据，这是共识。但何为高价值？却众说纷纭。

昆德拉认为：价值本身是一个基于主观的定义。同一组数据，在不同的时间、对于不同的对象，其价值是变化的。如果缺少背景，数据本身没有任何价值。丰田汽车召回事件发生之前，它的油门数据很少人会感兴趣。但事故一发生，就会有很多人需要它的数据。

昆德拉主张：凡是能增强公众对政府部门的问责能力和政府部门的反应速度、提高公众对于政府机关及其运作的了解、推进部门的核心使命、创造经济发展的机会或者满足公众特定要求的数据，就是高价值的数据。

6．网站虚实之争

Data.Gov的网站仅仅是一个数据的集散地，还是既提供数据又提供专门的分析工具？即便是集散地，也有虚实之分，是一个真正的数据仓库，还是一个数据索引库，源数据仍然保存在各个部门？

很多人认为Data.Gov应该是一个真正的数据集散地。奥巴马的公共支出开放网站、华盛顿的城际数据仓库都是真正的数据集散地。但昆德拉主张：联邦政府的部门太多，难以维护，Data.Gov网站应该是个索引库，各个部门仍然是各自数据的真正所有人和维护人。但用户从索引的链接跳跃到实体的数据，点击鼠标的次数不能超过3下。他还建议，网站建立初期，应该提供一些分析工具。

对每一个问题，昆德拉的心里其实都有自己的答案。但他知道，他还无法立刻说服大大小小的各级官僚。如果任由讨论继续下去，可能一两年也不会达成共识和实质性的结果。他坚信开放数据是正确的方向，他认为，如果一个社会真的要把开放作为一种正确的价值观来对待，就必须付诸行动。

他决定重点突破，先从一些没有争议的数据开始，并快刀斩乱麻，尽快推出一个技术平台。

2009年5月21日，距离奥巴马签署《透明和开放的政府》整整120天，Data.Gov上线发布了。

Data.Gov按原始数据、地理数据和数据工具三个门类组织开放的数据。上线的第一天，即使包括地理数据，这个新生网站上也仅仅只有47组数据、27个数据分析工具。

但即便如此，作为一个全国性的创举，Data.Gov还是受到了新闻界和大众的关注。上线第一天，该网站接受了210万的点击量，第二天又收获了250万的点击量。前两个月，创下了2000多万次的访问总量。

在众多的关注者中，有一家与众不同。他们从昆德拉一上任，就憋足了劲，全力支持Data.Gov，为它的每一个进步大声鼓呼。

这就是阳光基金会（Sunlight Foundation）。

就在Data.gov上线的同一天，5月21日，阳光基金会的主任艾伦·米勒（Ellen Miller）宣布设立25000美元的奖金，举办程序员公共数据开发大赛（App for America）。参赛作品必须使用Data.gov开放的公共数据，可以是一个数据分析的程序，也可以是一个数据可视化的应用，还可以是社交网站和智能手机的插件。

奥莱理出版社的CEO蒂姆·奥莱理又是本次大赛的赞助人和第一评委。

虽然Data.Gov初次上线只开放了47组数据，但3个月内，阳光基金会却收到了47个新开发的应用程序。这些程序，其中一些相当经典，以至于《纽约时报》、《华盛顿邮报》都对它们作了报道。当然，这些应用程序也成为昆德拉游说其他部门开放更多数据的超级“武器”。

8月26日，Data.Gov上一次性新增了178项原始数据。但昆德拉并没有松懈，他立足“互动”、不断完善Data.Gov平台的功能，先后加入了数

据的分级评定、高级搜索、用户交流以及和社交网站互动等等新的功能。例如，用户可以在网站上直接向联邦政府建议开放新的数据，而相关部门必须给出回应，若不同意开放，也要列出理由。从2009年5月至12月，Data.Gov共收到社会各界约900项开放数据的申请，联邦政府最后回复：16％的数据立即开放，26％将在短期内开放，36％将计划开放，还有22％因为国家安全、个人隐私以及技术方面的限制无法开放。

2009年12月8日，行政管理预算局（OMB）发布了《开放政府的指令》（Open Government Directive），命令各个联邦部门必须在45天之内、在Data.gov上至少再开放3项高价值的数据。

2010年2月，米勒又代表阳光基金会给昆德拉发出了一封公开信，对于政府各部门开放的高价值数据做了一次全面的评估，她直接批评11个机构对于开放数据的消极态度，并对数据的格式和质量提出了十分具体的建议。



迈克尔·克莱（Michael Klein）

迈克尔·克莱是一名退而不休的律师。2006年4月，他捐资600多万美元创立阳光基金会，该基金会的使命是“立足信息技术和互联网，促进政府的透明和开放”。



艾伦·米勒

阳光基金会的主要创建人之一。她已从事公益事业、草根运动35年，获得荣誉无数。还曾经创建过另外两家知名的公益机构：政治问责中心（Center for Responsive Politics）和公共运动（Public Campaign）。

随着这种从上到下、由外至里的高压态势，数据开放的步伐开始逐步加快。

2010年5月21日，Data.Gov上线发布的一周年纪念日，联邦政府开放数据的总数达到了27万项。

截至2011年12月，Data.Gov上共开放了原始数据3721项、地理数据386429项。

作为Data.Gov这艘旗舰的舰长，昆德拉再次获得了成功。他的成

功，也引起了美国各界甚至全世界的关注。昆德拉先后获得了一系列的奖项和认可，他被评选为年度CIO，被推选为世界经济论坛的青年领袖。

2011年，美国信息产业的巨头、全球500强之一的EMC公司，宣布建立“数据英雄奖”（Data Hero Award），以奖励那些“在大数据时代用数据对个人、组织、产业和世界产生了深远影响的从业人员”。2011年5月9日，EMC的评审委员会决定将首届“数据英雄奖”颁给昆德拉。舆论上下一致认为实至名归。

# 大众创新：航班延误之候机经济学

数据不会被它所激发的思想和创新消耗，相反，它可以为创新提供无穷的燃料。一小片合适的信息，可以促使创新迈进一大步。一组数据，可能会得到数据收集人难以想象的应用，也可能会在另一个看起来毫不相关的领域得到应用，因为这些创新型的应用，数据的能量将层层放大。[19]

——《利用数据的力量服务科学和社会》，联邦政府跨部门工作组给总统科学技术委员会的报告，2009年1月

保“数”运动（Save the Data Campaign）

Data.gov的发展并不是一帆风顺的。

2011年4月，由于预算赤字，美国联邦政府被迫大幅削减开支。国会宣布，原来用于支持Data.gov、USASpending.gov等数据开放网站的3400万美元年度预算将缩减到200万美元，这意味着个别网站可能被迫关闭。

以艾伦·米勒为首的阳光基金会立即联合13家公益组织，发起了“保数运动”。他们联合一些知名的技术精英和社会活动家，征集了3000多位公民的签名，给国会的众议长以及两党的领袖发出了公开信，要求国会在预算博弈中保护Data.gov。这场运动也受到国家公共电台等十多家媒体的报道和关注。

次月，国会表态，该项目的预算将最少增加到800万美元。

由政府主导、向全社会开放政府拥有的公共数据，这种做法，本身就是一种创新。但更重要的是，Data.Gov不仅仅是一个创新的结果，它的出现，代表着数据在社会的自由流动、知识向大众的自由流动，这为更多的大众创新、社会创新提供了一个平台。

截至2011年12月，仅仅在Data.Gov网站上，就汇集了1140个应用程序和软件工具、85个手机应用插件。其中，有近300个是由民间的程序员、公益组织等社会力量自发开发的。

本书将从中选择几个，立足细节，对美国社会如何利用公共数据进行创新做一个分析和介绍。

## 候机经济学：航班延误分析系统

近年来，因航班延误产生的机场纠纷一度在中国成为一个热点话题。

美国的航班也有延误，但却几乎没有“罢乘”、“霸机”、冲击机场的事件。Data.gov上线以后，美国交通部开放了全美航班起飞、到达、延误的数据，有程序员立刻利用这些数据开发了一个航班延误时间的分析系统（Flyontime.us）。该系统向全社会免费开放，任何人都可以通过它查询分析全国各次航班的延误率及机场等候时间。

这个系统上线之后，由于其简单、实用，获得了全美多个新闻报刊的报道和关注，成为很多人乘机、候机的行动指南。

大众创新和社会创新

大众创新，指以普通公民而不是以知识精英为主体的创新。

社会创新（Social Innovation），是指为了解决某个社会问题、满足某种社会需要、改善某部分人群的生存状况，民间力量自发产生的一种新的思想、行动和举措。社会创新也指政府在公共政策、社会治理方面的创新。

社会创新最著名的例子是孟加拉的乡村银行（Grameen Bank），它向贫穷的人发放不需要担保的小额贷款，以帮助穷人改善生活。其创建人Muhammad Yunus获得了2006年的诺贝尔和平奖。

以波士顿至纽约的航线为例（起飞：Boston，Logan International；到达：New York，Kennedy International），在系统的主页上，输入机场名称，点击之后，用户可以看到不同天气、不同日期、不同时段、不同航空公司、不同航班等各种条件下飞机是否准时以及平均延误时间的数据明细。

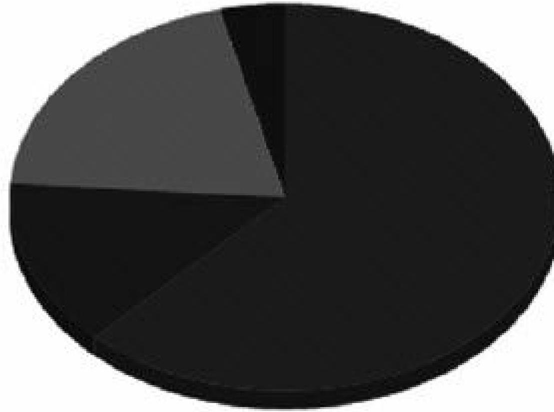
从以上分析可以看到，各种不同组合条件下的数据分析明细基本都有了。这些数据和分析结果，对具体一名消费者和整个社会的经济活动，会起到什么作用呢？

一是可以帮助消费者找到表现最佳、或者最符合自己需要的航班。如果没有这些信息，消费者在选择航空公司的时候，信息是不完全、不充分的，与航空公司构成一种典型的信息不对称关系。航班的历史数据是一种有效的参考和信号，公开这些信息，弥补了消费者的信息不对称。此外，消费者在对比分析大量历史数据的基础上，自己做出判断，即使结果不尽如人意，但也感觉公平。

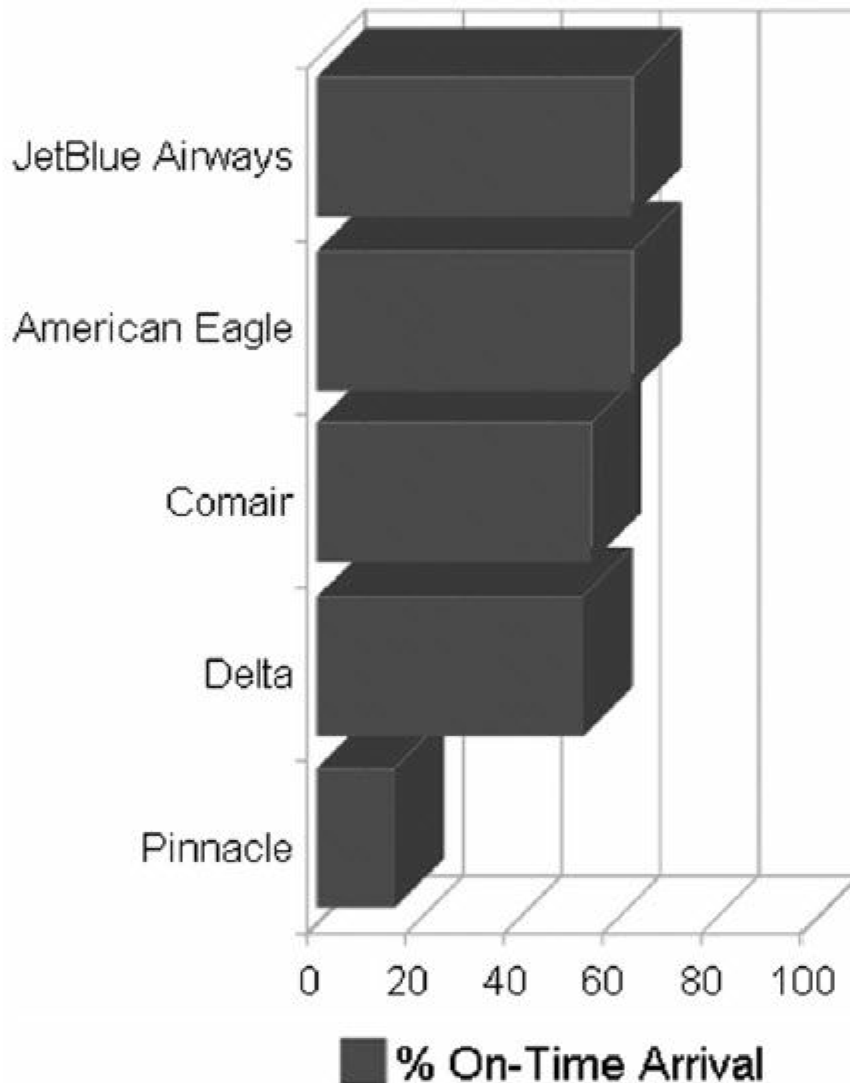
**Flyontime.us**的界面



总体情况分析


Flight Delay Summary
Based on 6735 flights from
01/01/2010 to 12/31/2010
62% On Time
14% 5-20 min. Delay
20% >20 min. Delay
4% Cancelled/Diverted

按航空公司分析




分析[1]：2010年从波士顿至纽约的航线总共有6735次航班，其中62％准点或提前到达，14％有20分钟以内的延误，20％有20分钟以上的延误，还有4％的航班最终取消。

分析[2]：按航空公司排序，JetBlue公司的准点率最高，为64.4％；Pinnacle公司的最低，为15.6％。

各种不同天气情况下的延误情况分析

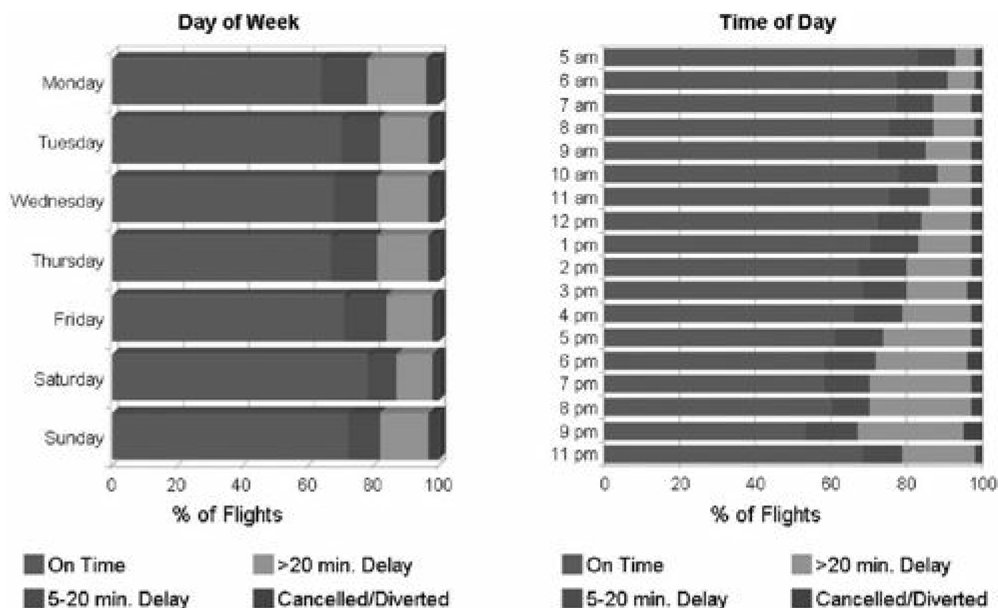
| | Average (median) | Be Prepared For (85th percentile) | Cancelled or diverted |
|---|---|---|---|
| In Good Weather... | 5 min. early | 19 min. late | 1% |
| In Bad Weather... | 3 min. late | 53 min. late | 7% |
| In Fog... (7%) | 12 min. late | 87 min. late | 14% |
| In Rain... (36%) | 4 min. late | 54 min. late | 7% |
| In Snow... (10%) | 7 min. late | 53 min. late | 12% |
| In Thunder... (5%) | 2 min. late | 53 min. late | 4% |

"Be Prepared For" gives the longest delay you can reasonably expect to occur. Only the unluckiest 15 percent of flights experience longer delays.

Current weather is 25.0 F (-3.9 C) and A Few Clouds, 39% humidity, and 10.00 miles visibility at BOSTON LOGAN INT'L ARPT, US, MA as of Sun, 18 Dec 2011 16:54:00 -0500.

分析[3]：天气良好的情况下，多数航班会提前5分钟到达；下雪天，平均延误时间为7分钟，年度最糟情况为53分钟，并有7%的航班取消；雷雨天气，平均延误2分钟，最糟情况53分钟。查询的结果还向用户显示当天当地的天气情况。

按星期和每天的时间段分析



分析[4]：就一星期而言，星期六的准点率最高，为78％；星期一的准点率最低，为64％。就一天而言，每天早晨5点到6点起飞的航班准点率最高，为83％；晚上9点到10点的准点率最低，为53％。

按节假日分析

| 假期 | 平均延误时间 | 最糟情况 | 航班取消 |
| --- | --- | --- | --- |
| Most Days | 2 min. early | 31 min. late | 4% |
| Day Before Memorial Day (239 flights) | 17 min. early | 3 min. late | 0% |
| Memorial Day (310 flights) | 6 min. early | 37 min. late | 1% |
| Day After Memorial Day (328 flights) | 4 min. late | 57 min. late | 7% |
| Labor Day (314 flights) | 5 min. early | 7 min. late | 2% |
| Wednesday Before Thanksgiving (353 flights) | 1 min. early | 26 min. late | 4% |
| Thanksgiving Day (214 flights) | 9 min. early | 6 min. late | 0% |
| Friday After Thanksgiving Day (247 flights) | 1 min. late | 21 min. late | 0% |
| Saturday After Thanksgiving Day (275 flights) | 5 min. early | 8 min. late | 0% |
| Sunday After Thanksgiving Day (334 flights) | 10 min. early | 10 min. late | 0% |
| Day Before Christmas Day (256 flights) | 14 min. early | 1 min. late | 1% |
| Christmas Day (232 flights) | 7 min. early | 17 min. late | 3% |
| Day After Christmas Day (283 flights) | 34 min. late | 80 min. late | 41% |

分析[5]：感恩节、圣诞节当天的飞机一般都会准点到达，但节日前后的延误情况是最严重的。圣诞节次日平均延误34分钟，最糟的情况是80分钟，还有41％的航班被取消。

按航班分析

| 航班号 | 平均延误时间 |
| --- | --- |
| Comair 6633 (10 flights) | 15 min. early |
| Comair 6663 (15 flights) | 10 min. early |
| American Eagle 4627 (182 flights) | 10 min. early |
| American Eagle 4377 (183 flights) | 9 min. early |
| Delta 575 (91 flights) | 9 min. early |
| JetBlue Airways 1015 (292 flights) | 8 min. early |
| **American Eagle 4614 (182 flights)** | **8 min. early** |
| JetBlue Airways 1005 (328 flights) | 7 min. early |
| JetBlue Airways 1003 (268 flights) | 7 min. early |
| …… | …… |
| JetBlue Airways 1013 (157 flights) | 3 min. early |
| American Eagle 4389 (183 flights) | 1 min. early |
| JetBlue Airways 1007 (249 flights) | 1 min. early |
| American Eagle 4643 (87 flights) | 1 min. early |
| Comair 6461 (114 flights) | on time |
| Comair 6257 (122 flights) | on time |
| American Eagle 4619 (182 flights) | on time |
| …… | …… |
| Delta 579 (4 flights) | on time |
| Comair 6467 (103 flights) | 1 min. late |
| …… | …… |
| Comair 6521 (110 flights) | 5 min. late |
| **American Eagle 4617 (182 flights)** | **7 min. late** |
| JetBlue Airways 1017 (363 flights) | 8 min. late |
| …… | …… |
| Comair 6345 (89 flights) | 22 min. late |
| Delta 1807 (40 flights) | 26 min. late |

分析[6]：全部航班的表现应有尽有。最准点的是航班Comair 6633，最不准点的是航班Delta 1807，平均延误26分钟。

二是最大程度降低了旅客等待时间的不确定性。法国有句谚语：让别人等待的人，他的错误会遭到众人的清算。[20]憎恶等待，是人之常情，因为等待意味着时间流失、经济损失，不确定性的等待还往往导致精神焦虑。单次航班的延误时间似乎是随机的、无规律的，但是，当数

据累积到一定程度时，航班延误时间的长短就会在统计上呈现出一种秩序和稳定。航班延误分析系统把这种统计学上“秩序和稳定”传达给了旅客，帮助他们建立正确的期待，合理安排时间，避免焦虑。

三是有利于推动航空市场的良性竞争。航班延误分析系统按平均延误时间给相关航空公司排了“座次”。回到上面的例子，经营波士顿至纽约航线的公司共有5个。就是否准点而言，谁好谁差，几乎一目了然。此外，各次航班的表现也有明细。例如American Eagle航空公司的第4617航班，全年共有182班次，平均延误7分钟；相比之下，该公司的4614航班，全年也是182班次，但平均提前8分钟到达。这些数据，不仅是消费者的行动指南，也是各大航空公司的核心竞争指标。通过这种数据公开，无疑可以促进市场竞争，航班延迟必然逐渐下降到消费者能够接受的合理范围之内。

其实，为了缓解航班延误的问题，美国政府也想过同样的办法。早在Data.Gov之前，国家交通安全局就在其网页上（NSA.gov）提供过一个“航班等待时间计算器”，帮助旅客估计因航班延误而导致的等待时间。2009年政府开放原始数据之后，民间开发出来的这个免费工具明显比交通安全局提供的“计算器”功能更强大、界面更友好。很快，该局便关闭了这个“计算器”。

有评论说，开放数据是一石三鸟，不仅服务大众，刺激经济，还调动了大众创新，为政府节省了软件开发的开支。

Flyontime.us还能够查询各个机场安检通关的时间长短，这个数据，当然是机场服务质量的一个重要指标。但这部分的数据来源，并不是政府发布的数据，而是乘客自己提交的数据。候机的乘客可以通过推特（Twitter）或者智能手机向该系统提交其在某个机场通过安全检查的

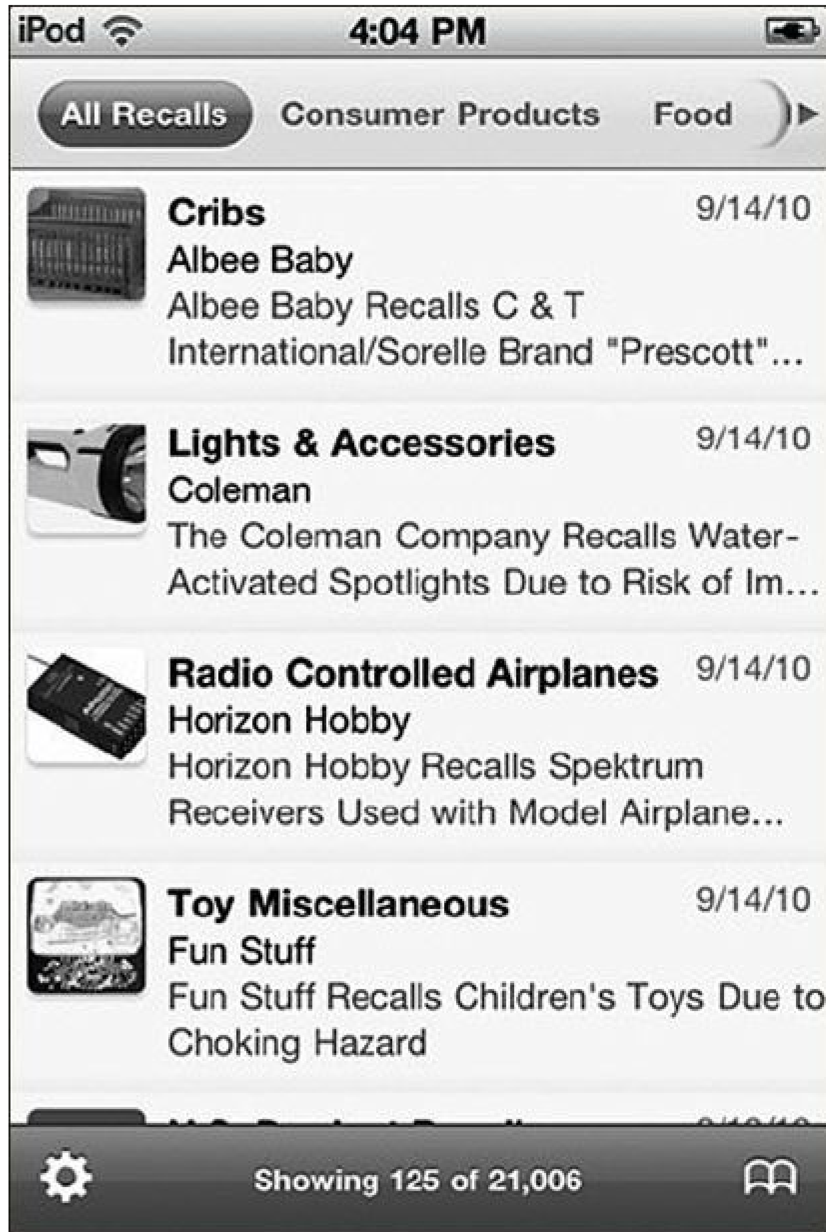
时间。这些数据，通过加总和平均，成为其他用户的参考。

当然，这又是一个大众通过网络的共同协作，改善管理、促进社会福利的例子。

有意思的是，伴随着Data.Gov的开放，美国的航班延误率正在呈下降趋势，由2008年的27％下降到2009年的20.79％，再到2010年的20.23％。当然，原因可能是多方面的，数据开放在其中的作用大小，还有待进一步研究。




联邦政府的**Recalls.Gov**手机界面




民间开发的**Recalls TM**手机界面（带有商品的照片和图像）

## 消费者的福音：商品召回手机查询系统

如何避免购买到具有安全隐患和质量问题的商品，无疑是所有消费者最关心的话题，这在美国也不例外。特别是婴幼儿的食品和用品，如果有质量问题，往往成为父母亲的梦魇。

Data.gov上线之后，各种因为质量问题而被召回的商品记录立即成为最受欢迎的数据之一。

联邦政府共有农业部、食品药品监督局、消费者委员会等6个部门涉及商品质量和产品召回的管理工作。早期，这6个部门在各自网页上公开其管辖范围之内的问题商品的召回记录。

2009年，为了方便用户查询，联邦政府内务部决定：在整合这些数据的基础上，开设一个专门的商品召回查询网站。在这个网站上，消费者可以对任何商品进行查询，获得其是否具有召回的记录，也可以对某一项商品进行跟踪，获得电子邮件的提示信息。

这个系统还可以在智能手机上使用。消费者在购物时，通过在随身携带的智能手机上键入产品的名称及型号，可以在购买时直接查看该商品是否具有质量问题和被召回的历史记录。如果消费者怀疑该商品存在问题或者是被召回产品的漏网之鱼，还可以立即在线举报。

但联邦政府开发的这个系统仅仅只能在Android（安卓）的手机上使用。

Data.gov开放数据之后，一款基于苹果手机的新应用Recalls™很快就被开发出来，并对全社会免费开放使用。和联邦政府开发的系统相比，该系统的功能更强大、使用更方便、界面也更绚丽，受到了消费者的莫大欢迎。

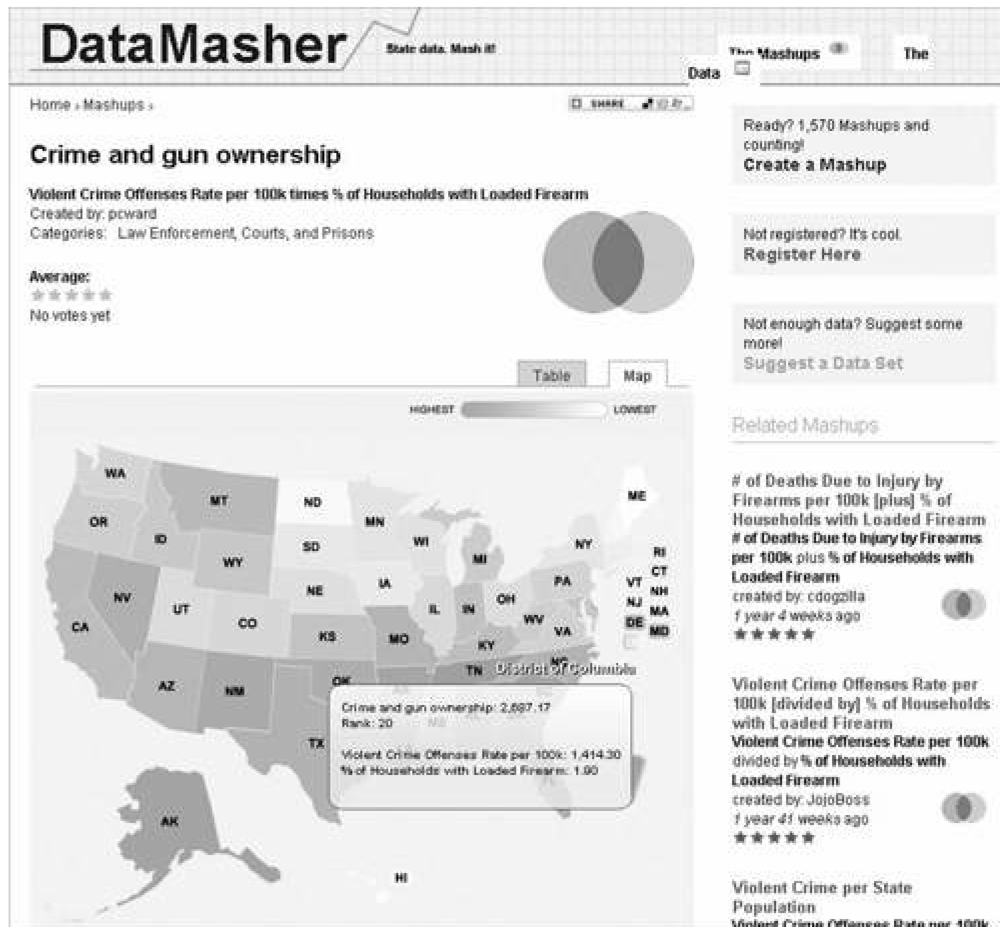
## 可视化的经典：数据混搭器

数据混搭器（Datamasher.org）是阳光基金会举办首次公共数据程序员开放大赛（App for America）的金奖作品。它之所以在众多的参赛

作品中拔得头筹，原因在于，它不仅仅是一个应用程序，还是一个平台，任何一个用户都可以在它上面选择需要整合对比的公共数据组，然后产生以地图为基础的数据可视化分析。

例如，一名用户以全美每个州为基准，整合了每百万人口发生犯罪案件的多少和各州居民拥有武器的比例两组数据，这两组数据分别来自联邦调查局和美国疾病控制中心（CDC）。下图表明，华盛顿居民的枪支拥有率为1.9％，远远低于全国平均水平，但其每百万人口的罪案数却为1414宗，居全国第一。又如密西西比州，其枪支拥有率为11.1％，居全国第二，但其每百万人口的罪案数却仅为291宗，在全国处于较低的水平。

犯罪案件的多少和居民枪支拥有率的可视化展示




这个可视化应用以直观的地图对比，想向读者证明：一个地区居民拥有枪支的多少，与该地区的治安情况没有必然的联系。

用户不仅可以在该平台上整合、分析Data.gov上开放的数据，还可以保存自己的设计和对比，并开放给平台的其他用户，和他人共享，接受他人的评价、意见和反馈。

截至2011年年底，该网站上已经拥有1570个不同的数据可视化应用。从一个地区空气质量与工厂数量多少的关系，到高中升学率与师资力量的关系，再到肥胖人员的数量与一个地区快餐店多少的关系，林林总总的对比和展示，有些给人直接的启示，有些发人思考，还有的令人忍俊不禁，但笑过之后又不禁感叹：社会大众蕴藏的智慧和创造力是无

穷无尽的！

## 注释

01 英语原文为："Every motivation that makes a man do something can be classified under'survival','social life'and'entertainment'. As a result, progress is defined as reaching a higher category: not doing a thing merely for survival, but for social reasons, and then, even better, just for fun."—*The Hacker Ethic and the Spirit of the Information Age*, Prologue, Linus Torvalds, 2001
02 英语原文为："The open society is one in which men have learned to be to some extent critical of taboos, and to base decisions on the authority of their own intelligence."—*The Open Society and Its Enemies*（1945）, Vol. 1, Endnotes to the Chapters: Notes to the Introduction
03 英语原文为："We're entering a new world in which data may be more important than software."—Tim O'Reilly
04 英语原文为："Who can afford to do professional work for nothing? What hobbyist can put 3-man years into programming, finding all bugs, documenting his product and distribute for free?"—An Open Letter to Hobbyists, Bill Gate, Feb 3rd, 1976
05 黑客（Hacker），其本意是指热衷、喜爱计算机编程、对程序设计有精深理解和高级技巧的人。
06 英语原文为："Given enough eyeballs, all bugs are shallow."—*The Cathedral and the Bazaar, Eric Raymond*
07 英文书名为*Open Sources: Voices from the Open Source Revolution*。
08 Fallibility, Wise Men and Politics, *New York Times*, May 20, 2008
09 8 Principles of Open Government Data, http://www.opengovdata.org
10 英语原文为："Information maintained by the Federal Government is a national asset."—Transparency and Open Government, Memorandum for the Heads of Executive Departments and Agencies, White House, 2009
11 英语原文为："Change will not come if we wait for some other person or some other time. We are the ones we've been waiting for. We are the change that we seek."—Barack Obama, speech, Feb. 5, 2008
12 该法案全名为：Federal Funding Accountability and Transparency Act of 2006。
13 OMB Offers an Easy Way to Follow the Money, Elizabeth Williamson, *The Washington Post*. December 13, 2007
14 参见2007年11月奥巴马访问谷歌的实况录像：Candidates@Google: Barack Obama, YouTube。
15 英语原文为："We are democratizing data, putting the power of information in the hands of the American people."—Vivek Kundra, January 22, 2010
16 华盛顿特区是美国联邦政府、国会和最高法院的所在地，其地位相当于中国的北京，在英文中简称为"Washington D.C."，这也是这个项目被简称为"DCStat"的原因。
17 President Obama Names Vivek Kundra Chief Information Officer, the White House, March 5, 2009
18 英语原文为："A primary goal of Data.gov is to improve access to Federal data and expand creative use of those data beyond the walls of government by encouraging innovative ideas（e.g., web applications）. Data.gov strives to make government more transparent and is committed to creating an unprecedented level of openness in Government. The openness derived from Data.gov will strengthen our Nation's democracy and promote efficiency and effectiveness in Government."参见Data.Gov网页。
19 英语原文为："Data are not consumed by the ideas and innovations they spark, but are an endless

fuel for creativity. A small bit of information, well found, can drive a giant leap of creativity. The power of a data set can be amplified by ingenuity through applications unimagined by the authors and distant from the original field."—*Harnessing the Power of Digital Data for Science and Society*, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, Jan 2009

[20] 法语原文为：On compte les défautsde qui se fait attendre.

# 第九章　试金石：白宫访客记录

你的所作所为和政府所做的一切事情同等重要。我请你追求一种公共利益，而不仅仅是自己的舒适生活；我请你捍卫我们必要的改革，使它免于失败；我请你从帮助你的邻居开始，服务这个国家。我请你做一个公民：一个公民——而不是旁观者，一个公民——而不是臣民，一个负责任的公民——建设我们服务型的社区和具有品质的国家。[01]

——乔治·布什，第43任美国总统，2001年1月20日就职演讲

作为开放政府的主要举措，Data.gov这艘旗舰启航之后，公共数据的无偿共享将数据开放的概念诠释得淋漓尽致，新闻舆论也给予了空前一致的好评。2009年，Data.gov的每一步推进，都获得了阵阵掌声。在一片赞誉和肯定声中，奥巴马要建立前所未有开放政府的雄心，似乎就要实现了。

但偏偏在这个节骨眼上，奥巴马自己却遭遇了一个直接来自民间草根组织的挑战。这个挑战，使他开放政府的雄心和诚意，受到了质疑。

故事缘于他主政之后面临的第一项重大历史任务：医疗体制改革。

# 总统在见谁：大医改中的“小”插曲

当我们放弃我们一直为之战斗的理念而妥协的时候，我们就失去了自我。在最困难之际，而不是最容易的时候，我们能坚守并高举这些理念，才能真正荣耀我们所推崇的东西。[02]

——奥巴马，接受诺贝尔和平奖时的演讲，2009年12月10日

医疗体制改革，早在20世纪初，就已经是美国政治版图中一个争议的焦点。

在所有发达国家当中，美国医疗系统的总体规模及人均开支都位居第一。以2009年的数据为例，全美的医疗支出占全国GDP的16％，比全世界的平均水平高出7个百分点；同时，全美人均医疗开支8000多美元，比平均工资的1/4还多，远远高出其他发达国家的相应比例。

但令人遗憾的是，花钱多仅仅意味着“贵”，并不等于“好”，更不代表“对”。几十年以来，医疗体制一直都是美国人屁股上的痛（Pain in Ass）。

原因在于，和其他国家相比，尽管美国在医疗领域的投入最大，但美国却是全世界发达国家当中唯一没有实现全民医保的国家，其医保覆盖率只有85％。全美约有4600万人生活在没有医疗保险的阴影之下。

没保险的人怨声载道可以理解，但拥有保险的人也不满意，因为金山银山并没有“堆”出质量的高山。2000年，在世界卫生组织（WHO）的排名榜上，美国医疗系统居全世界昂贵之首，但其绩效却只排全世界

第36名，美国人的总体健康水平在全世界191个国家当中只居第72位。[03]一些关键的医疗指标，如人均寿命，美国居全球第50位，又如婴儿死亡率，美国居全球第46位，都大大落后其他发达国家，与其超级大国的地位极不相称。

医疗，是基本的民生问题。可以想象，全民医保、医疗服务的质量，这两个问题也就一直都是美国立法者、执政者、朝野上下经年累月的热点话题。

其中，全民医保，又是其“结”中“结”。

作为平民政党，美国的民主党一直旗帜鲜明地支持全民医保，有很多位民主党总统，例如罗斯福、杜鲁门、肯尼迪、克林顿，都曾经扛起过全民医保改革的大旗，并为此殚精竭虑，但最终都在国会功败垂成。半个多世纪以来，全民医保始终是一盘没有下完的棋。

其中的根本原因，还是因为缺乏厚实的民意基础。美国人的传统，还是推崇“自助”，相信“自助是最好的帮助”[04]，而不是“他人的帮助”、“政府的帮助”或“上天的帮助”。本杰明·富兰克林（Benjamin Franklin）是美国的建国之父之一，他是一位发明家、文学家、政治家，平生拥有头衔无数，被誉为美国历史上“最完美的人”，他曾经对这个问题专门发表过见解：

“我认为帮助穷人最好的方法，不是给予金钱实物减轻他们的贫困，而是引导鼓励他们脱离贫穷。年轻的时候我在很多地方游历，在不同的国家我都观察到：政府为穷人提供的更多，他们自己努力的就更少，自己不努力，自然更加贫穷。相反，政府为他们提供的越少，他们就为自己做得越多，他们也就变得更富有。”[05]

也就是说，当一个国家的政府为人民负起太多责任的时候，人民就会失去“自助”的动力，不再对自己负责，社会的发展就可能会因此减速甚至停滞。

上百年来，富兰克林“自助自立”的观点，获得了多数美国人的认可，成为美国社会反对全民医保、反对建设福利国家的主要理由。

全民医保通不过，还有另外一个重要的文化原因：美国人认为行政权力不仅低效，而且很容易被滥用，因此从来就不信任政府。怀疑政府、限制政府，是美国人面对政府权力扩张的一种自然反应。全民医保，意味着政府要对医疗产业进行大规模的干预和包揽，这当然需要警惕。

全民医保的争论，也起源于《信息自由法》之争的摩斯时代。

和《信息自由法》过不去的约翰逊总统，却是全民医保的热心推进者，并在这个领域取得了历史性的成就。1965年7月4日，他在家中黯然签署了《信息自由法》，但没过几天，他就在白宫举行了隆重的仪式，在一片掌声中签署了《医疗保障法》（Medicare and Medicaid）。这个法案规定，由政府主导，为65岁以上的老人、残障人士和贫困人口提供医疗保险。这为美国的全民医保迈出了关键的第一步。

《医疗保障法》的立法史，甚至比《信息自由法》更难产，命运也更加多舛。

把全民医保正式提上立法议程的是第34任总统杜鲁门。

和摩斯一样，杜鲁门也是一位没有上过大学的民主党人。他务过农、当过兵、做过小生意，参加民主党之后，通过竞选担任了一个县的

法官，后来又当选为联邦参议员。1944年，他成为罗斯福竞选总统的搭档。次年，罗斯福在任内病逝，他由副总统接任总统，随后在任内大力推行“公平良政”（Fair Deal），全民医保就是其主要政纲之一。

也正是杜鲁门，在他的任内大兴冷战思维，在联邦政府解雇了2000多名同情共产主义的公务员，这成为摩斯提出《信息自由法》的直接诱因。1953年，摩斯提出《信息自由法》的时候，共和党正控制着国会，摩斯寸步难行。由于同样的党派之争，杜鲁门提出的“公平良政”计划也大多流产。

杜鲁门折戟之后，民主党的肯尼迪在举国欢呼中走进了白宫。他也反对《信息自由法》，支持全民医保。这位美国历史上最年轻的总统，颇能审时度势，他发现“全民医保”的目标并没有代表“全民民意”，就把受益人口缩小到老、弱、残等特定的人口群。肯尼迪虽然魅力四射、人气极高，但这个修改过的法案，在国会还是没有通过。

1963年，约翰逊上台，他提出了“大社会”（Great Society）的施政纲领，在民间激起了极大的反响。但他在全民医保方面的突破，与其说是“人力”，不如归为天时。因为这时候参加过二战的退伍老兵，开始步入中老年，医保成为一个很现实的问题。在美国，退伍军人是个极有分量的院外游说群体，有形容说：“退伍军人跺跺脚，国会山上抖三抖”。

约翰逊得到了退伍军人的支持，在他们的力挺之下，《医疗保障法》终于在1965年7月得以通过。

这时候的杜鲁门，已经是83岁的耄耋老人。



**1965**年**7**月，约翰逊总统举行隆重的仪式，签署了《医疗保障法》。为了纪念前总统杜鲁门对全民医保的开山之功，约翰逊在签署法案的同时，把第一张老人医疗保险卡授予了杜鲁门夫妇（图右部）。（图片来源：维基百科）

当我们放大历史上的种种细节，把很多件事情的前因后果联系起来，就很容易发现：历史的每一步发展，都相当的错综曲折。其中做出重大贡献的英雄人物，根本没有谁总是“伟大、光荣、正确”。历届美国总统，既是能干人，更是普通人，他们在以一己之力推动社会进步、国家富强的同时，无法面面俱到，甚至犯下了不少错误。但因为存在一个比较完善的制衡制度，这些错误，可以被纠正，也往往成为另外一些人建功立业的契机。美国的这些历史也证明，圣人治国是一个无法企及的良好愿望，甚至可以说，根本就是一个无稽之谈。

2009年，这个契机又临到了奥巴马。作为民主党人，奥巴马坚定地接过了“全民医保”这个火炬，从竞选总统开始，他就将扩大全民医保的

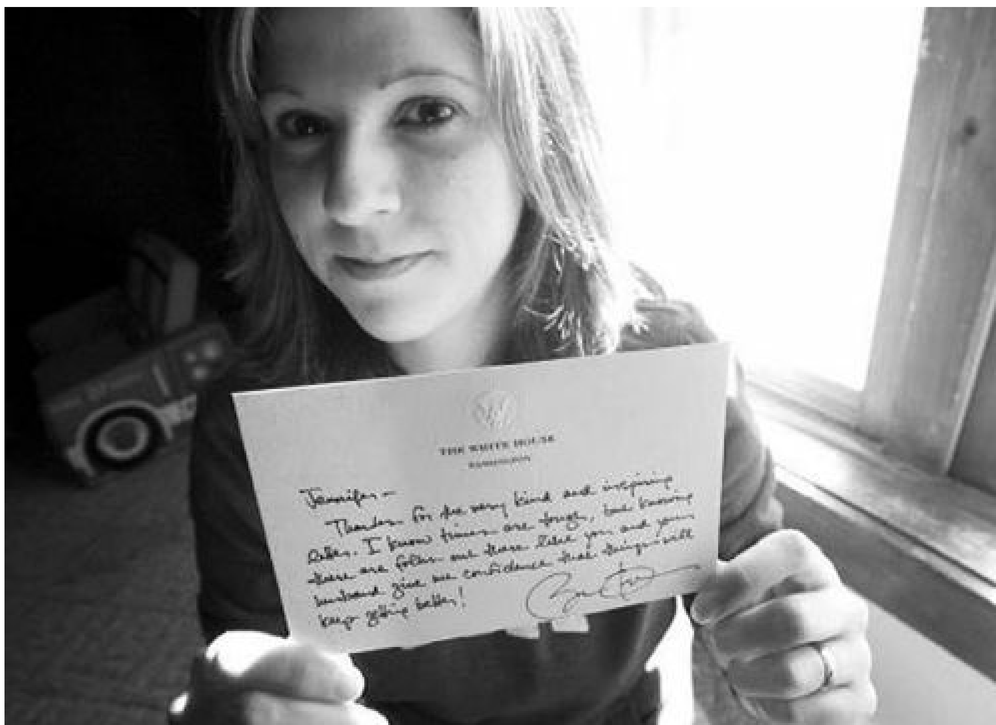
受益范围视为义不容辞的责任。

2009年3月，他一上任，就在白宫召开了医疗改革专项会议，重新拉开了这场改革的序幕。其后整整一年，为了推动这次改革的成功，他顶住了巨大的压力，使出了浑身解数。

奥巴马最初的计划是加大国家的财政投入，通过建立公营性质的医疗保险，扩大医保的覆盖面。但如何加大财政投入呢？他提出要对富人增税。这个方案几乎受到了共和党全党的反对，而且，由于强烈的行政干预色彩，绝大多数的美国民众也对此忐忑不安。

开局受挫之后，奥巴马没有轻言放弃。他深知其中的困难，也明白问题的关键在于民意。从上任的第一周起，他坚持每天晚上都阅读10封普通公民的来信。[06]这些信件当中，有不少声泪俱下地倾诉失业、生病、由于缺乏医疗保险带来的种种不幸。在医疗改革法案的演讲和辩论当中，奥巴马常常引用甚至背诵这些来信的内容，来支持他的政策和观点。

奥巴马多次引用过2010年1月8日密歇根州居民克莱因（Jennifer Cline）的来信。她在信中陈述了夫妻双双失业之后、发现自己身患癌症的痛苦和无助。奥巴马给她回信说：




克莱因向记者展示奥巴马的来信。由于这封信在医疗体制改革中的特殊地位，次年，她成功拍卖了这封信，用所得的**7000**美元偿还她的医疗欠款。（图片来源：网络）

“詹妮弗，谢谢你这封亲切和令人激励的来信。我知道时下生活艰辛，但是了解到你和你丈夫这样的人和事，给了我信心，我相信情况会不断好转！”

除了回信，奥巴马还在全国范围内争取普通人的支持。11岁的小男孩欧文斯（Marcelas Owens）是西雅图的一位中学生；2006年，他的母亲在失业之后，因为没有医疗保险、得不到有效的治疗而病故。欧文斯多次应邀来到白宫，在全国的新闻镜头前讲述他的亲身经历，他的话——“我不希望有任何孩子再经历我们家所遭受的痛苦”，打动了无数的美国大众。



**2010**年**3**月**21**日，听到医保法案在众议院通过的消息，奥巴马和**11**岁的男孩欧文斯击掌庆祝。（图片来源：**seattlepi.com**，**JEWEL SAMAD/AFP/Getty Images**）

2010年3月21日，美国众议院终于以219票对212票的微弱优势通过了医保法案，这是一项历史性的突破，美国的医疗保险覆盖率将从85％上升到95％，距离全民医保只有一步之遥！

但这个法案，也是一个妥协的产物。

奥巴马放弃了对富人增税的计划。最终的方案加强了对保险行业的管制，规定不容许保险公司因为受保人的病史、年龄等原因拒保或擅自增加保费。也就是说，负担的一部分转移到了保险公司。

这个重要的转变，发生在2009年5月。当时，美国的参众两院正在对奥巴马的医改方案展开激烈的辩论，民间的各种情绪非常高涨，反对的声音也愈发响亮。可以想象，当改革和争议的焦点转移到医疗保险界之后，为了自保，医疗保险行业对国会和白宫展开了密集的公关和游说。

一时间，全国的民众、媒体和各类公益组织都将关注的焦点集聚在保险行业的大公司、国会议员和奥巴马的一言一行、一举一动之上。

“华盛顿责任道德公民中心”（CREW）是一个位于华盛顿的公益组织。它致力于通过法律行动监督政府官员，防止公职人员为了特殊集团的利益而牺牲公共利益。

CREW的创始人是一位毕业于芝加哥大学的法律女博士：梅拉尼·斯隆（Melanie Sloan）。她曾经担任联邦政府华盛顿地区的助理检察官。2003年，她辞去公职，创办了CREW。其后一年半，整个组织只有她一个人孤军奋战。经过6年的打拼，2009年，CREW已经拥有了一支17个人的工作团队。

作为一个公益监督组织，CREW每年都组织评选首都地区最腐败的官员，并发布《最“腐败”报告》（CREW's Most Corrupt）。2006年，由于这份报告的揭露，资深众议员迪莱（Tom DeLay）被拉下马来，因选举洗钱罪获刑3年。

当政策争议的焦点转移到保险行业的时候，斯隆提出，奥巴马应该自证清白，公布自己在上任之后，与医疗保险行业的董事长、经理人以及相关说客在白宫的会面次数、时间长短和会议议题等记录。

斯隆说，人民有权知道，奥巴马在和谁见面、谁在访问总统？

这个17人的小团队，在全国的激辩和喧嚣中，发出了自己的强音：

人民有权知道——谁在访问白宫？谁在影响政策的制定？

在跌宕起伏的医疗体制改革当中，CREW的质询，只能算个小小的插曲。但这个质询，直接指向奥巴马。作为总统，奥巴马虽然在下一盘很大的棋，但对这个声音，他无法忽视。因为它有理有据，这样的声音，在一个民主社会，如果不正面回应，就会被新闻界不断放大，形成一股排山倒海的舆论压力。

# 全体美国人的房子：白宫

我从未忘记，我住在一幢属于全体美国人的房子里，他们将信任给予了我。我告诫自己，始终要记得，他们最深层次的问题是日常生活的问题。我常常倾听各种各样人的观点。他们当中，有这个国家产业界和金融机构的管理者，也有工人和农民的代表，但最多的，还是一些并非身居高位的普通公民。[07]

——富兰克林·罗斯福，第32任美国总统，1938年4月14日

白宫，位于华盛顿特区宾夕法尼亚西北大道1600号，是美国总统居住和办公的地点。白宫始建于美国的第一位总统华盛顿，但由于工期过长，他任内无福消受。等到1800年，美国的第二任总统亚当斯（John Adams）才正式搬进白宫。

从此，美国各任总统及其家人都在这里办公、居住。

白宫最初只是被笼统地称为总统官邸。1901年，因为其白色的外表，第26任总统西奥多·罗斯福（Theodore Roosevelt）将其形象地称为“White House”，此后一直沿用。“White House”如果直译，是“白色的房子”，而不是“白色的宫殿”（White Palace），把“房子”硬生生地译成“宫”，大概是中国人的皇权意识在作祟。

罗斯福之后，“白房子”就慢慢演变为美国最高决策机构的代名词。

“白房子”其实并不大，包括地下室、洗手间，才上下6层、167个房间。[08]其产权全部属于国家，这也是美国仅有的一座由纳税人支付的个

人住宅。所以准确地说，全体美国人都是美国第一家庭的房东。也正因为如此，多年来，历届总统都主动将白房子称为“人民的房子”（People's House）。



白宫外观，可能还不如很多中国的市政府气派。（图片来源：网络）

上个世纪30年代，美国进入大萧条时期。为了增强美国人民的信心，民意大师罗斯福总统（Franklin D. Roosevelt）利用刚刚兴起的广播技术，和美国人民在电台上展开了著名的“炉边谈话”（Fireside chat）。1938年4月14日晚，他就说道：“我从未忘记，我住在一幢属于全体美国人的房子里，他们将信任给予了我”。

2008年8月2日，奥巴马在竞选中接受全国城市联盟（NUL）的采访时，也亲口说过：如果当选，“白房子”将是人民的房子，欢迎大家来参观访问。

正因为产权的所有人、房子的使用人都认为这是“人民的房子”，白

宫有非常悠久的对外开放历史。自美国的第三任总统杰斐逊起，白宫就对外开放，公民可以在不影响总统办公的前提下参观白宫。杰斐逊本人甚至会时不时走出办公室，和参观的游人握手寒暄、喝茶聊天。当然，当时的华盛顿，地广人稀、交通不便，并不是每天都有人来探访总统。现在的白宫，已不可同日而语。虽然还是有新年、国庆开放日，并且部分区域在规定时间内也向游客开放，但要见一面美国总统，需要通过层层的申请和检查，绝非易事。

白宫的安保和门卫由国土安全部下属的特勤局（USSS）负责，所有人员进出白宫，无论公务还是私事，都要由特勤局录入一个专门的数据库，该数据记录被称为“白宫访客记录”（Visitor Log）。

白宫的布局和设计是公开的，美国有不少网站提供互动的白宫结构演示图




白宫1楼西翼的房间布局。白宫共有167个房间（其中有35个是洗手

间）、3座电梯。白宫中间部分为住宅区，椭圆形办公室位于白宫1楼西翼。（图片来源：MSN网站白宫内部结构互动程序截图）




2012年4月，白宫和谷歌合作，将内部房间的实景进行360度立体拍摄，并放置在网上。（图为谷歌地图网站上白宫内部的截图）

早在布什时代，华盛顿责任道德公民中心（CREW　　　）就曾经因为“白宫访客记录”，与总统布什结下了“梁子”。

2006年，CREW曾经根据《信息自由法》，要求特勤局公开布什总统和几位宗教领袖会面的数据记录，但遭到布什政府的拒绝。布什认为，这是总统记录，而不是联邦政府的行政记录，总统见什么人、谈什么话、谈了多久，都不在《信息自由法》的监督管辖范围之内。

CREW随即将特勤局告上了联邦法院。

2007年12月，联邦地方法院宣判，认定“白宫访客记录”属于《信息自由法》规定的行政信息，白宫必须公开。[09]

布什不服，向联邦巡回法院提出了上诉。2009年1月20日，奥巴马

入主白宫的时候，该上诉案还在审理当中。

奥巴马入主白宫之后，其实也采取了不少亲民措施。和布什相比，他确实加大了白宫的开放力度，他在上任第一天就开放了白宫，接见了上百名来白宫参观的普通公民。这在近年来并不多见，前一次是1993年，克林顿上任的第一天，曾用抽奖的形式邀请了2000名普通公民访问白宫。



**2009**年**5**月**8**日，奥巴马邀请部分白宫工作人员的家庭成员访问白宫。其中一位小男孩非常好奇奥巴马的头发是不是和他的一样，奥巴马弯腰低头让他抚摸。（摄影**: Pete Souza**）

2009年4月，奥巴马上任不久，CREW又依据《信息自由法》向奥巴马政府提交了一则信息公开的要求，要求公开两名“煤老板”访问白宫的记录。这时候的白宫，虽然换了主人，但巡回法院的上诉还没有结果，因此仍然保持了布什时代的口径，强调“白宫访客记录”不属于行政公开的范围，拒绝了CREW的要求。

CREW当然不甘示弱。2009年6月再次将奥巴马治下的特勤局告上了法庭。

这时候，美国的参众两院正在对医疗体制改革进行激辩，全国上下群情沸腾、议论纷纷。6月22日，CREW又提出，为了分析确定医疗行业、保险行业的大公司对于奥巴马制定医保政策的影响，白宫应该公开奥巴马上任之后和这些行业的董事长、经理人会面的次数和记录。

CREW一口气开出了一份列有18位医疗、保险行业顶级经理人的名单，又根据《信息自由法》，向白宫提交了公开奥巴马和他们会面情况的申请。

白宫再次拒绝，CREW又吃了一次“闭门羹”。

面对新闻界，斯隆公开批评奥巴马说：“他没有遵守参选时许下的诺言，现在的白宫，并不是人民的白宫！他也没有落实他就任时的承诺，这样的政府，既不开放，也不透明！”

这时候，“白宫访客记录”，俨然已经成了一块检验奥巴马是否真正开放的试金石。

# 步步妥协：总统与草根的对决

防止公民犯下错误，这并不是我们政府的责任；但防止政府陷于错误，却恰恰正是每一个公民的责任。[10]

——罗伯特·杰克逊（1892-1954），美国最高法院大法官、司法部长

在斯隆强烈要求奥巴马公布商界巨子访问白宫及与其会面记录的时候，轰轰烈烈的医疗体制改革，已经攻守易势。此时的奥巴马，由改革的发起人，变成了改革的辩护人，面对压倒性的反对意见，正举步维艰。

2009年6月底、7月初，各种大小民意调查都表明，奥巴马的支持率在一路下滑。皮尤研究中心（PRC）的民调表明，他的个人支持率由上任初期的64％下滑至7月份的54％；短短3个月内，其医改计划的支持率也从51％下滑至42％，已经不足半数。[11]

《赫芬顿邮报》（*The Huffington Post*）是美国著名的政治新闻媒体之一。关注政治，当然要关心民调。2008年该集团成立了Pollster.com网站，专门致力于“民调中的民调”。所谓“民调中的民调”，是指在收集所有民调公司对某一特定问题的调查结果之后，根据它们的样本大小和调查方法，对它们的调查结果进行加总和分析，最后形成自己的观点和看法。这种结果，当然比一个调查公司发布的数据更加全面可靠。

《赫芬顿邮报》将数十家专业调查机构对奥巴马支持率和反对率的调查结果绘成散点图，再根据这些点的分布确定其支持率和反对率的变

化曲线。结果表明，正是在推行医疗改革法案的一年期间，奥巴马的支持率迅猛下降。

面对前人无法完成、盘根错节的医疗体制改革，即使支持率大幅下降，奥巴马还是鼓足勇气、迎难而上。但要公开自己和老板、说客的会面记录，他还是不免反复掂量、犹豫再三。

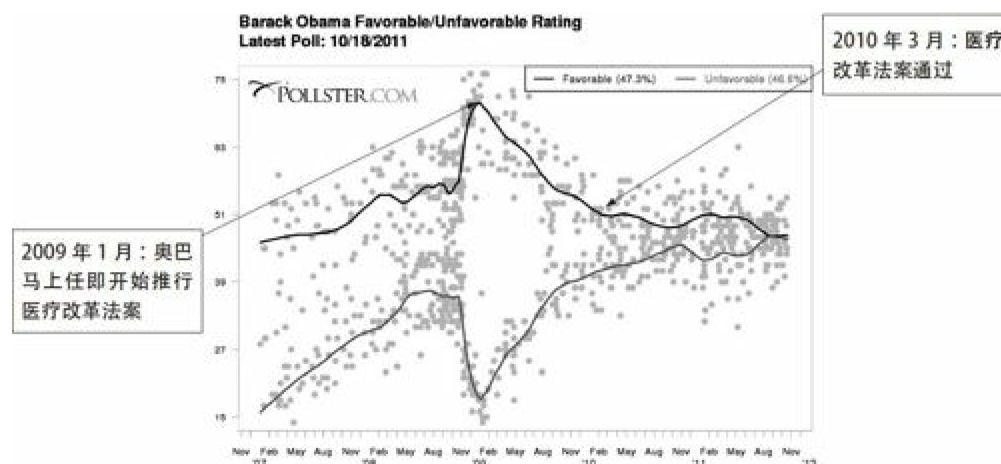
从4月底到6月底，短短两个月，CREW已经两度将白宫特勤局告上法庭，白宫却岿然不动，奥巴马也没有任何表态。

随着医疗体制改革的步步深入，奥巴马的形势变得更加孤立、艰难，白宫的态度也开始发生微妙的变化。

7月7日，CREW意外收到白宫的一封来信。这一次，白宫没有再提“拒绝”二字。而且，信中还解释说，尽管白宫在联邦巡回法院的上诉还没有结果，但已经开始自行检讨现行的政策是否合理。

这是一个僵局解冻的信号？

民调中的民调：推行医疗改革法案期间奥巴马的支持率大幅下降



说明：上方深色线为支持率的变化曲线，周围的点为各个民调机构调查得出的支持率；下方浅色线为反对率的变化曲线，周围的点标示反对率的分布。

白宫的主动示好，很快就有了答案和解释。

7月22日晚，为了争取民众和舆论的支持，奥巴马召开了医疗改革新闻发布会。

就在这次发布会开始前的几个小时，CREW突然又收到了一封白宫来信，这一次，还是总统顾问克雷格（Greg Craig）的特快专递。

出人意料的是，克雷格在这封信中一股脑儿公开了18位“敏感人士”在奥巴马上任之后访问白宫的日期和次数。

但令人意外的是，对于白宫的妥协，CREW并不买账。

第二天，斯隆接受了媒体采访。她在电台中指出，这封两张纸的回信只是一个情况的总结，并没有完整地回应CREW的信息公开要求。斯隆再次向新闻界说明了CREW的诉求：

白宫

2009年7月22日立即发布

尊敬的魏斯曼女士[12]：

为了回复你在2009年6月22日、根据《信息自由法》给美国国土安全部提交的信息公开要求，我特此给你回信。你要求披露自2009年1月

21日以来18位医疗商业领域高级管理人员到访白宫的记录。我知道，美国国土安全部已在2009年7月7日回复过你的要求。该回复指出，关于白宫访客记录的案件还在诉讼当中，白宫也正在检讨我们的政策，考察是否能够无条件地开放这些记录。

医疗体制改革是当前最重要的问题。总统正全力以赴，力争让美国人过上更健康的生活，提高国家的竞争力。鉴于公众对医疗法案辩论的强烈兴趣，以及总统设定的要增加政府透明度的目标，针对你提交的18位人员的名单，总统决定行使其自由裁量权，我们查阅了截至2009年6月30日的白宫访客记录，现公布以下记录：

•Bill Tauzin曾于3月5日、5月19日、6月2日及6月24日访问过白宫。

•Karen Ignagni曾于3月5日、6日、11日及6月30日访问过白宫。

•Richard Umbdenstock曾于2月4日、23日，3月5日、25日、30日，4月6日及5月22日访问过白宫。

•J. James Rohack曾于3月25日、6月22日及6月24日访问过白宫。

•William C. Weldon曾于5月12日访问过白宫。

•Jeffrey B. Kindler曾于3月5日、5月6日及6月2日访问过白宫。

•Stephen J. Hemsley曾于3月15日及22日访问过白宫。

•Angela F. Braly曾于2月13日访问过白宫。

•George Halvorson曾于3月27日及6月5日访问过白宫。

•Jay Gellert曾于2月10日、3月11日及20日访问过白宫。

•Thomas Priselac曾于4月3日访问过白宫。

•Richard Clark曾于3月24日访问过白宫。

•Wayne T. Smith曾于6月4日访问过白宫。

•Rick Smith曾于5月19日及6月2日访问过白宫。

除此之外，白宫访问记录表明：Bill Tauzin、Karen Ignagni、Richard Umbdenstock、J. James Rohack、Jeffrey B. Kindler、George Halvorson、Jay Gellert和Thomas Priselac都曾计划参加5月11日在白宫召开的会议。除了Jeffrey B. Kindler以外，其他人全部到会。最后，以下人员没有访问白宫的记录：Ben J. Lipps，William A. Hawkins II，Robert L. Parkinson。

我们将会继续审阅你根据《信息自由法》提出的信息公开要求和开放白宫访客记录的政策。

诚挚的

格雷戈·克雷格（总统顾问）

“我们需要的，是原始的数据。从原始的数据记录当中，我们可以推断出更多的情况。白宫如果不公布全部的数据，我们就无法知道这封信是不是有所隐瞒、是否列出了全部的访客记录。”

斯隆继续批评说，这封来信，只对CREW关于医疗保险行业的信息

公开要求作出了简单的回应，对另外几次信息公开要求根本只字未提。很明显，这是一种敷衍和搪塞，是奥巴马为了当晚的新闻发布会能够顺利召开、避免在大庭广众之下产生不和谐声音的缓兵之计。

斯隆的判断是对的。

在当晚新闻发布会的现场，奥巴马确实被新闻记者“将”了一军。《芝加哥论坛报》的新闻记者帕森（Christi Parsons，*Chicago Tribune*）向他直面发问：

“在你竞选期间，你承诺将在C-SPAN公共电视台上直播关于医疗体制改革的会谈。但是，白宫最近拒绝公开你和一些医疗保险行业高级经理人会面的记录。此外，救市计划的总督察官也披露说白宫阻留了很多救市的信息。我的问题是：你有没有兑现关于建立一个透明政府的承诺？”

这个问题咄咄逼人，但因为有备在先，奥巴马得以一言化解、轻松“脱身”。他回答说，他相信他的这届政府确实要比往届更为开放透明。他和这些经理人见面的时候，大部分时间都有记者在现场拍照。对于CREW信息公开的要求，他也已经做出了正面的回应。

但他所谓的“正面”回应明显不能令人满意，斯隆第二天公开拒绝了奥巴马的解释。她在接受采访时表示，白宫需要公开的是全部访客的原始数据记录。只要一天不公开，CREW就会坚持诉讼的手段，战斗到底。

这一次，奥巴马没有挺过两个月。

2009年9月4日，他对此发表了正式的总统声明，作出了真正的“正

面回应”。

这是一个全面的妥协和让步。

奥巴马在声明中说：

“从本月起，‘白宫访客记录’将在互联网上公开发布。这是美国历史上首次公开这些记录。我们不仅要把白宫的大门向美国人民打开，还要让阳光照在白宫正在发生的事情之上。这样，本届政府才能实现建立历史上最开放、最透明政府的目标。美国人民有权利知道我们在政策制定的过程中正在听取谁的声音。”[13]

根据奥巴马的声明，白宫出台了相应的“白宫公开访客记录的自愿政策”[14]。该政策表示，白宫将放弃正在法院审理之中的上诉，将从2009年9月起，公开发布不涉及国家安全和机密的因公到访白宫的记录。每个月公布90天到120天之前的数据记录，对于需要保密的记录，白宫不公开其内容，但会公布保密记录的数量和多少。

当天，CREW又收到了总统特别顾问艾森（Norm Eisen）的来信。艾森在信中代表奥巴马，向CREW为促进这项新政策的诞生所做出的努力致以“诚挚的”谢意。

斯隆发表声明说，奥巴马兑现了自己的承诺，这是历史性的进步。CREW非常骄傲能在这个历史性的进步当中作出自己的贡献。



梅拉尼·斯隆

华盛顿责任道德公民中心（CREW）的主任。她深信“正是一小批执著的公民改变了世界”，并将这句话写进了该中心的使命陈述。2008年，她被《滚石》杂志评为“引发改变的100人”（100 Agents of Change）之一。

斯隆的故事被广泛报道。也是这个9月份，她被著名的女性杂志《欧普拉》（*The Oprah Magazine*）[15]评选为美国20名最具有力量的女性之一。在采访中，她谈到了自己不懈“战斗”的动机和初衷：

“当我每天打开报纸、读到那些负面新闻的时候，我就想：这太糟糕了！为什么没有人去做点什么改变它呢？我不仅仅是想，我放下报纸，这就去做！”

CREW的抗争画上了一个美丽的句号。

但“白宫访客记录”的故事，却还没有结束。新的质疑、新的问题还在继续出现。

# 从白宫安保到政治监督：执著的公民改变世界

永远不要怀疑：那一小部分有思想并且执著努力的公民能够改变这个世界。事实上，人类的历史从来都是这样。[16]

——玛格丽特·米德（1901-1978），美国文化人类学家

奥巴马的妥协并没有获得全社会百分之百的认可和肯定。

前文提到过的阳光基金会主任米勒在9月4日当天就提出：每个月只发布90天到120天之前的数据，是一个巨大的政策瑕疵。白宫在发布数据之前，确实需要一些时间来对数据进行预处理，例如清除数据中的个人电话号码、个人社会安全号码（SSN）等隐私信息。但现代的信息技术已经允许我们大大缩短这个周期，21世纪的政府应该以实时的形式发布数据。

米勒呼吁，白宫公开数据的周期应该缩短到一天，每天都应该发布前一天的访客记录。

此外，阳光基金会还组织技术人员，把白宫发布的数据记录和其他数据库进行了增值整合。通过到访人员的姓名、白宫访问记录和游说人员的捐款数据库[17]、知名人物社会关系数据库（LittleSis.org）[18]，以及维基（Wikipedia）、谷歌搜索（Google）连接整合起来，让所有游说人员到访白宫的次数、捐款的多少以及各类社会关系等相关信息都在一个页面中一览无余。虽然可能因为人员的名字相同，导致张冠李戴，但这种“全景式”的浏览还是给监督机构和研究人员提供了巨大的方便。[19]

白宫访客记录的数据也很快就成为Data.Gov的一部分。没过多久，2010年4月，又有人利用这些数据在Data.Gov上开发了一个新的查询工具：白宫访客搜索（White House Visitor Search），这个工具，不仅可以按访问人、被访问人对原始数据进行查询，还利用语义网的技术，把相关人员的数据与其他网站的联接起来。

另一家公益组织“司法监督”（Judicial Watch）的主任认为，白宫应该公布所有的访客信息，除了因公的访问，因私访问总统和副总统的记录，也应该公开发布。他讽刺说：除了他们不想公开的，白宫确实公开了“所有的”信息。

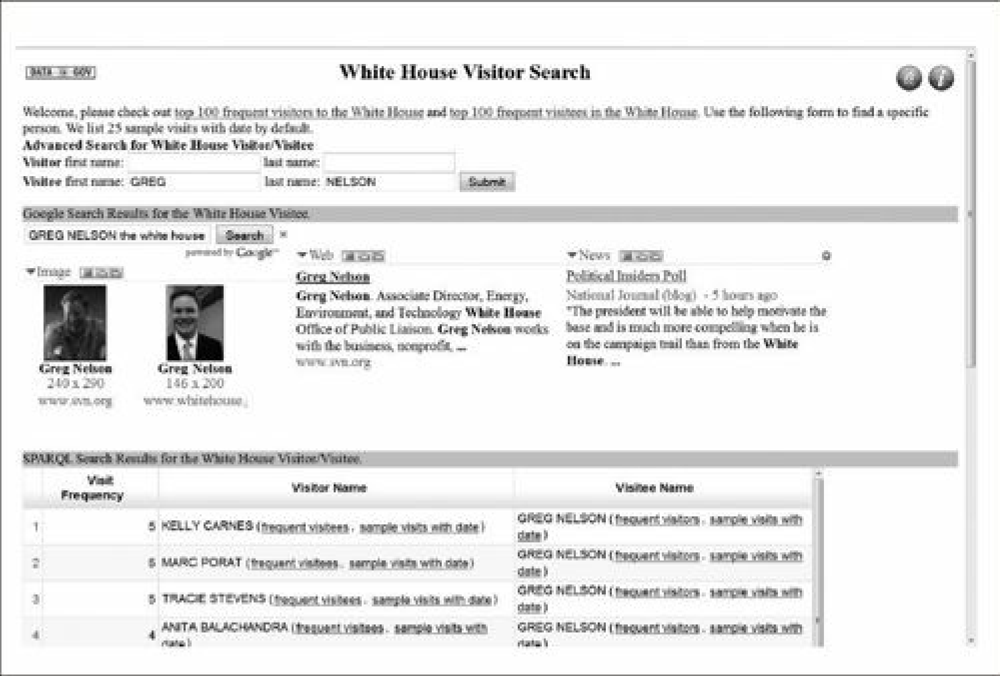
一波未平，一波又起。

2011年4月，“白宫访客记录”又引发了新一轮的批评。公共廉政中心（CPI），是一家致力于推进政府透明的新闻调查机构，该机构在研究了白宫从2009年9月到2011年4月公布的100多万条数据记录之后，得出了以下结论：

白宫公开的“白宫访客记录”数据字段说明和实例

| 数据字段 | 中文解释 | 数据实例 |
| --- | --- | --- |
| NAMELAST | 来访人的姓氏 | FREELY |
| NAMEFIRST | 来访人的名字 | JOSHUA |
| NAMEMID | 来访人的中间名 | L |
| UIN | 会谈序列号 | U59226 |
| BDGNBR | 出入标识卡号码 | 75885 |
| ACCESS_TYPE | 进入白宫的方式 | VA |
| TOA | 到达白宫的时间 | 11/17/2010 8:59 |
| POA | 进入白宫的入口岗位 | D1S01 |
| TOD | 离开白宫的时间 | 11/17/2010 13:56 |
| POD | 离开白宫的出口岗位 | D1S01 |
| APPT_MADE_DATE | 定约日期 | 11/15/2010 17:52 |
| APPT_START_DATE | 会谈开始时间 | 11/17/2010 9:30 |
| APPT_END_DATE | 会谈结束时间 | 11/17/2010 23:59 |
| APPT_CANCEL_DATE | 会谈是否取消 | |
| TOTAL_PEOPLE | 总共来访人数 | 1 |
| LAST_UPDATEDBY | 数据更新人 | YT |
| POST | 数据输入的计算机标识 | WIN |
| LAST_ENTRY_DATE | 最后修改日期 | 11/15/2010 17:52 |
| TERMINAL_SUFFIX | 数据输录人 | YT |
| VISITEE_NAMELAST | 被访问人的姓 | METZENBAUM |
| VISITEE_NAMEFIRST | 被访问人的名字 | SHELLEY（OMB 的一位副主任） |
| MEETING_LOC | 会谈所在的大楼 | WH |
| MEETING_ROOM | 会谈所在的房间 | WW |
| CALLER_NAME_LAST | 提交免除审查申请人员的姓 | ALTOFT |
| CALLER_NAME_FIRST | 提交免除审查申请人员的名字 | LOIS |
| DESCRIPTION | 事件描述 | COMING IN FOR MEETING W/ S.METZENBAUM AND C.MUNOZ |
| RELEASE_DATE | 数据发布日期 | 2/25/2011 |

白宫访客搜索（**White House Visitor Search**）的界面




说明：这是一个语义网的新应用，把访问人、被访问人的数据和其他的数据自动联接起来，方便用户的查询。本书后续将介绍语义网的技术。

1．有人去了白宫，但白宫公布的数据当中却没有相应的记录；

2．白宫办公厅主任的受访记录仅有500余次，但他的3个助手却共有4400次；

3．有205000个记录的事件描述是空白的，其中，很多都是和总统及其高级助手召开的小型会议，会议目的是什么，无从得知。[20]

此外，还有报纸报道，白宫现在经常把和游说人员见面的会议安排在白宫旁边的附属办公楼，如此这般，白宫的访客记录中就能不留痕迹。很明显，通过把被访问人的姓名记成低级助手、转移会议地点，个别高级官员就能逃避公众的监督。

2011年5月3日，在一次国会的听证会中，共和党的众议员斯特恩斯（Cliff Stearns）援引公共廉政中心发布的研究结果，批评奥巴马的民主党政府没有信守“透明行政”的承诺。

白宫的发言人立即对此作出了回应。他们解释说，白宫的助手常常是会议和访谈的第一联系人和协调人，他们名字出现的次数较多是正常的。白宫认为，“白宫访客记录”之所以多次招致批评，最根本的原因在于，当初开发设计这个系统，是为了维持白宫的秩序、保护总统和副总统的安全，而不是政治监督。

这个回答，确实也道出了问题的关键和白宫的苦衷。

但有理由相信，一个新的、更实用的白宫访客记录系统，已经在一些公益组织和执著公民的酝酿当中。美国的白宫，不久后将会安装上一套适合政治监督的访客信息管理系统。

因为，历史的发展从来都是这样——执著的公民改变世界。

对白宫访客记录穷追猛打的**4**个公益组织

CREW citizens for responsibility and ethics in washington

名称：华盛顿责任和道德公民中心

使命：通过有影响力的法律行为，打击为了特殊利益而牺牲公共利益的官员，建设一个更美好的华盛顿。

Sunlight Foundation

名称：阳光基金会

使命：使用最先进的信息技术来督促政府变得更透明和负责。

The Center for Public Integrity iwatch news Investigation. Impact. Integrity.

名称：公共廉政中心

使命：揭露权力滥用、腐败和失职，使政府运行更加诚实、正直和负责，并把公共利益放在第一。

Judicial Watch Because no one is above the law!

名称：司法监督

使命：在政府、政治和法律领域推广正直、透明和责任，因为没有人高过法律！

## 注释

01 英语原文为："What you do is as important as anything government does. I ask you to seek a common good beyond your comfort; to defend needed reforms against easy attacks; to serve your nation, beginning with your neighbor. I ask you to be citizens: citizens, not spectators; citizens, not subjects; responsible citizens, building communities of service and a nation of character."—George W. Bush, First Inaugural Address, Saturday, January 20, 2001

02 英语原文为："We lose ourselves when we compromise the very ideals that we fight to defend. And we honor those ideals by upholding them not when it's easy, but when it is hard."—Barack Obama, Nobel Lecture, Dec. 10, 2009

03 The World Health Report 2000, World Health Organization

04 这是一句来自于伊索寓言的古谚语，其原文为：Self help is the best help. Aesop.

05 英语原文为："I think the best way of doing good to the poor, is not making them easy in poverty, but leading or driving them out of it. In my youth I traveled much, and I observed in different countries, that the more public provisions were made for the poor, the less they provided for themselves, and of course became poorer. And, on the contrary, the less was done for them, the more they did for themselves, and became richer."—Benjamin Franklin

06 奥巴马不仅读信，还亲笔回信；不仅读美国人的信，还读外国人的信。据报道，中国公民李良正、吴迪等人就曾先后收到过他的回信。

07 英语原文是："I never forget that I live in a house owned by all the American people and that I have been given their trust. I try always to remember that their deepest problems are human. I constantly talk with those who come to tell me their own points of view—with those who manage the great industries and financial institutions of the country—with those who represent the farmer and the worker—and often, very often with average citizens without high position who come to this house."—On Economic Conditions, Franklin Roosevelt, April 14, 1938

08 见白宫网页：http://www.whitehouse.gov/about/History/

09 White House Visitor Logs Are Public, Judge Rules, *New York Times*, December 18, 2007

10 英语原文为："It is not the function of our Government to keep the citizen from falling into error; it

is the function of the citizen to keep the Government from falling into error."—Robert H. Jackson

11 Obama's Ratings Slide across the Board, Pew Research Center, http://people-press.org/2009/07/30/obamas-ratings-slide-across-the-board/

12 魏斯曼是CREW的首席顾问，是信息公开要求的提交人和署名人。

13 英语原文是：For the first time in history, records of White House visitors will be made available to the public on an ongoing basis．We will achieve our goal of making this administration the most open and transparent administration in history not only by opening the doors of the White House to more Americans, but by shining a light on the business conducted inside it．Americans have a right to know whose voices are being heard in the policymaking process.

14 White House Voluntary Disclosure Policy Visitor Access Records, http://www.whitehouse.gov/VoluntaryDisclosure/

15 *The Oprah Magazine* 是著名脱口秀主持人欧普拉创办的杂志，由于她的知名度，该杂志在美国非常流行畅销。

16 英语原文为:"Never doubt that a small group of thoughtful, committed citizens can change the world. Indeed, it is the only thing that ever has."—Margaret Mead

17 OpenSecrets.org和TransparencyData.com是美国两个公益数据网站，免费提供数据库追踪联邦层面和州政府层面个人和组织对各类竞选的捐款。

18 这是一个致力于确定知名人物和各类社会组织之间关系的数据网站。截至2011年7月，该网站已经为60456名各类社会知名人物和18660个组织之间建立了319112条关系，其中包括游说者11117名。

19 详细说明可见：http://sunlightfoundation.com/WhiteHouseVisitors/

20 White House visitor logs riddled with holes, Center for Public Integrity, http://www.iwatchnews.org/2011/04/13/4115/white-house-visitor-logs-riddled-holes

# 第十章　矿难中的歌声和数据

从来就没有什么救世主，

也不靠神仙皇帝。

要创造人类的幸福，

全靠我们自己。[01]

——欧仁·鲍狄埃，《国际歌》，法国，1871年

2010年4月5日，奥巴马刚刚签署《医疗改革法案》不到半个月。他还沉浸在这个划时代历史成就的兴奋当中。

突然，西弗吉尼亚州传来一声巨响，美国发生了一起震惊全国的矿难。

这起矿难共导致29名矿工死亡，死亡人数之多，为近40年之最。

而且，又是西弗吉尼亚州。

100多年以来，作为美国最重要的产煤区，西弗州一直为美国源源不断地输送煤和能量，但西弗州同时也是著名的矿难高发地区，不断地制造各种悲情记录。1907年，该州曾经发生过美国历史上最惨烈的莫农加煤矿（Monongah）大爆炸，造成362人死亡。

多年来，这里的矿工，最牵动美国人的神经。

在美国，死亡人数超过5人的矿区事故，就被定义为“灾难”（disaster）。1976年以来，死亡5人以上的矿难仅仅发生过13起，平均每两年一起。

这一宗死亡29人的矿难，很快占据了全国媒体的头条和封面，不少报纸和杂志对其进行了长篇累牍地分析和报道。来自全国各地的电子邮件和信件像雪花一样涌进了白宫。

这些信，开头几乎都一样：“我很骄傲我来自一个矿工的家庭”、“我是一名矿工的儿子”、“作为一名矿工的女儿，我很自豪”。除了感到自豪之外，这些矿工无一例外，都要求奥巴马采取行动，查明原因。奥巴马对新闻界提到这些诉求，他说：“他们要求我关注矿工、为他们祷告。他们说，别忘了，是矿工维持着美国的光和亮。在信中，他们提出一个简单的要求：别让它再发生，别让同样的事再度发生！”

奥巴马还公开承诺：“一个依赖矿工的国家怎能不竭尽全力保护他们？我们的国家怎么能容忍人们仅仅为了工作、为了追求美国梦就付出生命的代价？”

4月25日，奥巴马参加了这次矿难的追悼会。在开始演讲前，他逐一念出了29名遇难矿工的名字。然后，他对矿工的生活进行了描绘：

“凌晨4点半，最晚5点，他们就起床，开始一天在黑暗中的工作。穿着厚重的工作服、硬头靴和安全帽，开始一小时的车程，他们静静地坐着，直到车开进群山之中、五英里深处的矿井。黑暗中，唯一的灯光是他们头上的安全帽和车辆的前灯。

日复一日，他们挖煤掘矿；年复一年，我们理所当然地享用着他们的劳动成果：这些能量，照亮了我们的会议中心，点亮了我们的教堂、

家园、学校和办公室，维持着我们国家和整个世界的运行。

很多时候，他们从黑暗的矿里探出头来，眯着眼睛盯着亮光。

很多时候，他们从肮脏的矿里钻出身来，满身汗水、尘土和煤灰。



**4**月**25**日，奥巴马在西弗吉尼亚州参加遇难矿工的追悼仪式，和遇难者的家属在一起。（图片来源：**Official White House, Photo by Pete Souza**）

大多数时候，他们可以回家，但这一天，他们没有！”[02]

奥巴马还提到，矿工是一个团结的群体，他们之间像家人一样彼此相待，很多人都为自己属于这个群体而感到骄傲。

奥巴马的描述，高度浓缩了一个多世纪以来，美国人对于矿工群体的共同印象：艰苦但友爱，悲情却自豪。

矿工群体的自豪，不仅仅是因为他们为美国的经济发展输送动力和燃料，更重要的是，这个群体在美国的劳工史上、民权史上，扮演过先锋的角色，起过不可替代的作用，留下了宝贵的遗产和深深的印记。

那是一段激荡、燃烧、斗争的岁月。

要理解美国人对于矿难和矿工的集体记忆，就必须先了解这一段历史。

# 集体行动的号角：你站在哪一边

每一个公民，都应该是一个战士。古希腊和古罗马就是这样，每一个自由的国家都必定如此。[03]

——托马斯·杰斐逊（1743-1826），第3任美国总统

美国曾经是矿难高发的国家。它的开矿史，可以说是一部矿工的“血泪抗争史”。

20世纪前30年，美国矿难不断，平均每年有2000名以上的矿工因事故死亡。1907年，死亡人数曾高达3442人，创下了历史之最。

像世界上其他的国家一样，当时美国矿难频发的根本原因，是矿区的管理者没有平衡好追求利润和安全生产的关系。100年前的美国煤矿，生产方式落后，联邦政府也没有系统的法律来监管矿区的安全，阴暗潮湿的工作环境、简陋陈旧的挖掘设备、松散的管理制度以及封闭的信息流通渠道，这些都是导致矿难发生的重要原因。

**1901**年至**2010**年美国矿难死亡人数的情况

| 年度 | 矿难死亡总人数 | 平均每年死亡人数 | 每万名矿工死亡率 |
|---|---|---|---|
| 1901-1910 | 22,739 | 2,274 | 36.7 |
| 1911-1920 | 24,680 | 2,468 | 32.9 |
| 1921-1930 | 22,252 | 2,225 | 29.4 |
| 1931-1940 | 12,528 | 1,253 | 22.5 |
| 1941-1950 | 10,907 | 1,091 | 22.3 |
| 1951-1960 | 4,512 | 451 | 15.7 |
| 1961-1970 | 2,597 | 260 | 17.5 |
| 1971-1980 | 1,420 | 142 | 6.8 |
| 1981-1990 | 877 | 88 | 4.5 |
| 1991-2000 | 426 | 43 | 3.2 |
| 2001-2010 | 328 | 33 | 2.7 |

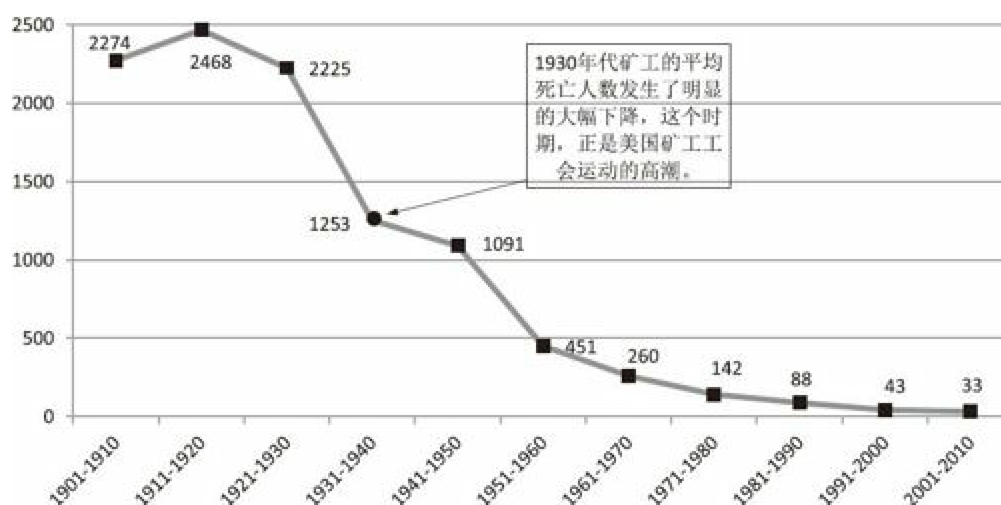
说明：本章数据只计算煤矿，不包括金属矿等其他矿藏。（数据来源：美国矿山安全健康局）

但从1930年起，美国矿难的死亡人数开始大幅下降。其直接原因，是联邦政府加强了监管、资本家更重视对安全的投入。但根本原因，还是因为矿工的“自助”——矿工通过组织工会、集体抗争，在20世纪上半叶掀起了一股劳工运动的高潮。

美国的煤矿工人工会，是美国劳工运动史上的先驱。矿工工会曾经为铁路、汽车、钢铁、纺织等各个行业的工会提供了宝贵的组织经验、输送了人才的血液，在美国的工会历史上有不可替代的地位。

美国联合矿工工会（UMWA）正式成立于1890年。但这时候的工会，虽然受到了矿工群体的认可和欢迎，却得不到资本家一方的承认，在法律上也没有正式的地位。在重大的问题上，资本家一方仍然我行我素，无视工会的存在。

**1901**年至**2010**年美国矿难年均死亡人数的变化趋势



（数据来源：美国矿山安全与健康局）

本书第六章曾经探讨过“公共利益缺乏代表的困局”和“集体行动的逻辑”。政治经济学家奥尔森指出，利益受到侵害的大众，由于其分散性，很难有效地组织起来；而公司、资本家一方，由于其利益高度集中，很容易结盟，形成游说的势力，左右政府的政策。

这也正是矿工和工会所面临的困局。工会的作用，是通过一个有效的中心，将矿工们组织、动员起来，通过谈判、罢工等“集体行动”保护群体的利益。但困局在于，这边你罢工，那边资本家可以另找其他人来开工。就是在同意罢工的队伍里面，也有投机取巧、冷眼旁观甚至见风使舵之人。而一旦罢工成功，所有的工人都将从中受益，但罢工行动的领导者、参与者，却可能要付出非同一般的个人代价。

这种代价，在1930年前后的美国，意味着可能要牺牲个人的生命和家庭的安全。

当时的资本家为了阻止矿工参加工会、瓦解工会组织的行动，无所不用其极：他们雇用打手维护矿区的秩序，驱赶工会的领袖，甚至派出枪手在工会领袖的家里投掷炸弹、杀人灭口，手段非常残酷。

而矿工工会则组织纠察队、拉警戒线，集聚在矿区的入口，试图阻止资本家雇用的新工人进入矿区上班。

在美国矿工工会的抗争史上，曾经发生过很多动人心弦的故事。这些故事，对整个美国的工会运动甚至后来的民权运动都产生了深远的影响。

1931年，肯塔基州哈兰县的矿区发生了罢工，资本家和工会之间发生了激烈的冲突。在这场冲突中，警察和政府站到了资本家的一边，想帮助他们尽快恢复生产和秩序。

一天晚上，一群煤矿主雇佣的私家打手在警察的带领下，气势汹汹冲进该地区工会领导人里斯（Sam Reece）的家里，对其进行抓捕。

里斯闻讯后仓促逃离，留下了一家妻小。荷枪实弹的警察和打手们在里斯家里翻箱倒柜、大声呵斥，希望找到一点蛛丝马迹。受到惊吓的一群孩子，在一旁放声大哭。里斯的妻子弗洛伦斯，也是一名矿工的女儿，她搂着自己的儿女、义正词严地质问警察说：

“你到底站在那一边？这里除了饥饿的孩子，什么也没有！”

一番肆虐之后，警察和打手扬长而去，留下了弗洛伦斯在一片狼藉的厨房里饮泣。这天深夜，她撕下了一张日历纸，咬着嘴唇、在泪水中写下了一首歌：《你站在哪一边？》。

这首歌套用了一首宗教赞美诗[04]的旋律，如泣如诉，反复吟唱一个问句：你站在哪一边？这像是质问，也像是倾诉，更像是呼唤。由于这首歌朗朗上口、极富感染力，很快就在矿区的工人当中流行起来。

当资本家组织另一批工人进入矿区、重新开工的时候，罢工的矿工则集聚在矿区的入口，他们对着准备迈进矿区大门的新工人唱起这首歌。

歌声呼吁矿工们站到一起，直到这场斗争赢得胜利；歌词质问道：你是要做一个为人不齿的帮凶，还是做一个真正的男人？歌中最后概括说：穷人除非组织起来，否则不会有任何的机会！

他们试图用歌声来打动、说服其他的工人参加他们的集体罢工行动。

奥尔森在指出集体行动的困境之后，也曾经为这种难以团结的局面开出了药方。他的办法是，一个群体，要建立合适的激励手段，营造为共同利益而努力的文化氛围和运行机制。毫无疑问，《你站在哪一边？》这首歌曲，就代表了这样一种文化。

类似的工会抗争一直持续到1933年6月。这当中，有不少工会的领袖都为抗争付出了鲜血的代价。在劳资关系急剧紧张的情况下，时任总统的罗斯福不得不采取对策。在他的推动下，美国国会以微弱的多数通过了《国家产业恢复法》（National Industrial Recovery Act），该法案在法律上明文赋予了矿工加入工会的自由以及工会代表矿工和资本家进行集体谈判的权利。

这部法律是美国劳工历史上的一个里程碑。

从此以后，矿工工会的活动取得了合法性。每一次矿难发生，都会引发工会的介入，而资本家、联邦政府和国会都不得不以最快的速度来应对工会可能采取的集体行动。结果就是，每一次事故，都会引起相关法律的更新和完善，在事故—立法、新事故—新立法的不断循环中，美国先后制定了十几部法律，安全生产的标准越来越高，生产事故也就越来越少。同时，因为工会的作用越来越大，矿工的工资福利也越来越好。

你站在哪一边？

词：弗洛伦斯·里斯

来吧，我的好兄弟，

我给你们带来了一个好消息：

咱们自己的工会，

已经在这儿建立。

合唱：

你站在哪一边，

告诉我，

你站在哪一边？

我爸爸是个矿工，

现在他已经不能下井，

跟他站在一起吧，

我的矿工兄弟，

直到这场斗争赢得胜利。

他们说：在哈兰县里，

没有中间立场。

你要么做一名工会的会员，

要么选择当布莱尔的帮凶。

噢，朋友，

你还能忍受吗？

噢，请你告诉我，

你要当一个为人不齿的帮凶，

还是做一个真正的男人？

不要当老板们的走狗，

不要听信他们的谎言。

除非我们穷人组织起来，

否则不会有任何一个机会。

Which Side Are You On?

Lyrics: Florence Reece

Come all of you good workers

Good news to you I'll tell,

Of how the good old union

Has come in here to dwell.

CHORUS:

Which side are you on,

Tell me,

which side are you on?

My daddy was a miner,

He's now in the air an' sun,

Stick with him,

Brother miners,

Until this battle's won.

They say in Harlan county,

here are no neutrals there,

You'll either be a union man,

Or a thug for J. H. Blair.

O gentlemen,

Can you stand it,

O tell me if you can,

Will you be a lousy scab,

Or will you be a man?

Don't scab for the bosses,

Don't listen to their lies.

Us poor folks haven't got a chance,

Unless we organize.



弗洛伦斯·里斯（1900-1986）

除了《你站在哪一边？》，弗洛伦斯还写过很多首劳工歌曲，并出版了诗歌集《逆流》。1981年，里根政府大幅削减社会福利计划，此时她虽然已是高龄老人，但还发表新诗对里根进行批评。（图片来源：The Civil Rights Digital Library，University of Georgia）

1951年，伊利诺伊州发生了一次矿难，119人丧生，这直接催生了1952年的《联邦煤矿安全法案》（Federal Coal Mine Safety Act of 1952）。1968年，西弗州的一次矿难导致了78名矿工的死亡，这时候的总统是憎恶《信息自由法》、支持全民医保的约翰逊，他立即做出快速反应，直接向国会递交了《联邦煤矿健康和安全法案》（Federal Coal Mine Health and Safety Act of 1969），其后通过了美国历史上最严格、最全面的煤矿安全法规。该法规定，全国所有的地下煤矿，每年必须接受联邦机构的4次检查，违规者将受到罚款和刑事起诉。再后来，联邦政府成立了独立的安全监察部门——矿山安全健康局（MSHA）。

弗洛伦斯也一直活跃在工会运动的一线，很多次受邀在罢工现场演唱《你站在哪一边？》这首歌曲，为工人的集体行动鼓气助阵。

《你站在哪一边？》不仅在矿工群体流行，在各行各业的罢工运动当中都被传唱。为了切应时境，它的歌词被改编成十多个版本，但不论如何修改，都大同小异，“你站在哪一边”这句质问和呼唤，始终都是这首歌的核心。

1960年代，美国开始了轰轰烈烈的民权运动，这首歌又被皮特·西格（Pete Seeger）、鲍勃·迪伦（Bob Dylan）等一批著名的歌手引用、翻唱，成为民权运动中呼吁“集体行动”的号角。这首歌也被输出到英

国、意大利、加拿大等很多个国家，成为这些国家罢工场合的“主旋律”。



1973年美国麻省的工人在罢工中学唱《你站在哪一边？》（图片来源：Radical Passions，Kendall Hale）

2011年10月，美国兴起了“占领华尔街”的运动，这场运动随后波及全球。很自然，在华尔街的队伍当中，同样的歌声再次响起。

## 可以避免的悲剧：数据揭示的全景式真相

政府数据作为一项公共资源，应该像天气预报、体育赛事和股票信息一样实时公开。通过把信息的力量放到民众的手中，可以增加公民对公共事务的参与、对政府的监督。

——维伟克·昆德拉，美国联邦政府第一任首席信息官，2010年

2010年4月5日西弗州发生爆炸的煤矿叫做上大支矿（Upper Big Branch Mine）。

上大支矿隶属于梅西公司（Massey Energy）。

梅西公司是全美第四大煤矿公司，拥有46口矿井，6000名矿工，控制着23亿吨的煤矿储备量，年产量高达4000万吨，堪称西弗州的主要经济支柱。可以想象，作为一个行业巨头，该公司在当地拥有广泛的影响力。

但上大支矿的所有矿工，却都不是美国联合矿工工会的会员。

矿难发生的第二天，美国联合矿工工会就对矿难发表了声明。工会的发言人很遗憾地指出：上大支矿的矿工，没有一个是工会的会员。工会的负责人还指出，上大支矿的爆炸，初步判断，是一场可以避免的悲剧，如果有工会的存在，这样的悲剧就绝不会上演！

但现实无法假设。

光阴荏苒，潮起潮落，曾经轰轰烈烈的美国联合煤矿工会，这时候已经光彩不在，颇有日落西山之感。

其中的原因，很简单、也很现实。

一方面，是随着生产率的提高和能源的多样化，美国社会对矿工的依赖已经大不如前。1923年，美国曾经有近90万名在井下作业的矿工，到2010年，矿工群体已经锐减到12.9万人，这个数字，还包括在矿区工作、并不下井的文秘人员。1960年以后，美国煤矿业开始大幅裁员，矿工相继失业，工会也无法保住他们的饭碗。

另一方面，随着制度的健全、生产条件的改善，煤矿行业的危险性已大大降低。2007年，煤矿业的事故死亡率仅为0.0277％，甚至低于农牧渔业的0.0279％。而矿工的收入，却比大部分行业都高。

这种情况下，越来越多的矿工觉得没有必要再加入工会。到2010年，美国联合矿工工会其实只覆盖了全国不到20％的在职矿工。

当然，具体到上大支矿，还有更复杂的原因。随着事故调查的深入，真相层层揭开：梅西公司的管理层曾不遗余力地阻止矿工加入工会，他们通过支付比平均工资还高的工资，换取了该矿矿工不加入工会的承诺。

这个消息经披露之后，当然令很多人都感到一种无言的沉重。

但真正牵动大众神经的，还是矿难发生的真正原因。这场被美国煤矿工会断言为“可以避免的悲剧”，到底是怎样酿成的呢？

像中国一样，矿难之后，美国的互联网上民情汹涌，声讨责任的声浪此起彼伏。美国网民也在网上挖掘灾难发生的原因，但与国内盛行的“人肉搜索”不同，随着公共数据的大面积开放，美国民众可以通过数据发现真相。

矿难发生的当天，民众舆论的矛头首先直指政府主管部门——美国矿山安全健康局，第一个问题是：该局是否工作到位、监管得当？

登录矿山安全健康局的网站，可以看到“煤矿警告、命令和安全”的数据库，输入“上大支矿”，返回逾千条数据条目，这些数据，是该局对上大支矿进行监管的流水记录。每一条记录，都包括检查的时间、结果、违反的法律条款、处理的意见、罚款的多少、已缴纳的金额、煤矿是否申诉等数据项。

这些数据表明，上大支矿的安全早已亮了红灯。该煤矿仅2010年3月份就有53条违规记录，其中12条与地下逃生通道和通风设施有关。2009年，上大支矿共有515条违规记录，被处以总计为897325美元的罚款，违规次数是全国煤矿平均水平的11.6倍。

2010年1月7日，灾难发生前的3个月，上大支矿又因为通风设施的问题，被矿山安全健康局处以66142美元罚款，是罚款金额最大的一次。

既然早已劣迹斑斑，为何迟迟没有关闭？——其实，这1000多条数据记录也提供了答案。矿山安全健康局的官员解释说：为了保证公平，安全检查中设有申诉制度，在最后的申诉结果出来之前，无法关闭煤矿。不幸的是，申诉过程复杂冗长，梅西公司正好以此作为缓兵之计，不断提出申诉，最终导致了悲剧的发生。数据显示，在202项比较严重的警告中，梅西公司对154项进行了申诉；对于前100宗大额罚款，梅西公司只认了15％的账。

矿山安全健康局网站上公布的对上大支矿监管的数据记录

**Citations, Orders, and Safeguards**
**Mine ID: 4608436 Mine Name: Upper Big Branch Mine**

I Indicates violations that have not yet been assessed.
II Indicates violations pending hearings, appeals, and/or other actions.
III These are non-assessable.

Note: Vacated citations are not included in any reports on the DRS.

| Violator | Contractor ID | Citation/ Order No. | Case No. | Date Issued | Final Order Date | Section of Act | Date Terminated | Citation/ Order | S & S | Standard | Proposed Penalty ($) | Citation/ Order Status | Current Penalty($) | Amount Paid To Date ($) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Massey Energy | I | 8098204 | 000222782 | 3/30/2010 | 7/21/2010 | 104(a) | 4/9/2010 | C | N | 75.370(a)(1) | 285.00 | Closed | 285.00 | 285.00 |
| Massey Energy | II | 8085078 | 000222782 | 3/25/2010 | | 104(a) | 3/30/2010 | C | Y | 72.630(b) | 2,106.00 | In Contest | 2,106.00 | 0.00 |
| Massey Energy | III | 8087763 | | 3/24/2010 | | 104(b) | 3/24/2010 | O | N/A | | Non-Assessable | | | |
| Massey Energy | II | 8086928 | 000222782 | 3/23/2010 | | 104(a) | 3/23/2010 | C | Y | 75.370(a)(1) | 3,405.00 | In Contest | 3,405.00 | 0.00 |
| Massey Energy | II | 8086927 | 000222782 | 3/23/2010 | | 104(a) | 3/23/2010 | C | N | 75.370(a)(1) | 263.00 | In Contest | 263.00 | 0.00 |
| Massey Energy | II | 8084612 | 000222782 | 3/23/2010 | | 104(a) | 3/23/2010 | C | N | 75.325(b) | 138.00 | In Contest | 138.00 | 0.00 |

4月6日，矿难发生的第二天，矿山安全健康局特事特办，将上大支矿的监管记录打包、专门陈列到一个文件中，以供网民浏览。这些数据是政府职能部门“有所作为”的最好注脚，矿山安全健康局也因此在民众的视线中迅速“脱责”，关注的焦点很快转移到梅西公司的管理层上。

政府一而再、再而三地警告处罚，这貌似一场完全可以避免的灾难，梅西公司的老板是何方神圣，胆敢如此漠视政府的三令五申、顶风作案呢？

这时候，Data.gov网站上已经公开了从地理人口到经济能源等几十万项来自政府各部门的数据资料，网民很快就在该网站上查询到了答案。梅西公司的老板叫布兰肯希普（Don Blankenship），其名下有46个煤矿，全部的名字、地点、所有人和经营人等资料在Data.gov上都条目清晰。

布兰肯希普的政治影响力也很快以数据的形式浮出了水面。

在一个叫“透明数据”（TransparencyData.com）的网站上，网民可以查询到和梅西公司相关的全部政治捐款记录。“透明数据”网站收录了全美联邦、州两级政治竞选的全部捐助情况，从竞选捐款人、受益人、捐款额，到时间、地点，甚至竞选结果的数据都在该网站上免费共享了。数据表明，自2003年以来，梅西公司的政治行动委员会有过326笔政治捐款共计287万美元，绝大部分是布兰肯希普个人捐出的。稍做数据分析，网民就发现了捐款的两个特点。一是近90％的捐款14笔共248.55万美元流向了一个叫“以孩子的名义”的组织；二是接受款项最多的个人是西弗州地方上诉法院的法官本杰明（Brent Benjamin），计24笔共2.25万美元，其次是现任州议员安第斯（Troy Andes）。

“以孩子的名义”也是一个竞选组织，它在新闻网页和视频网站上留下的唯一脚印是一场罢免地方上诉法庭麦格劳（Warren McGraw）法官的选举运动。顺藤摸瓜，网民很快发现，这位法官曾作出过一系列对梅西公司不利的判决，至于现任州议员安第斯，曾在梅西公司的公关部工作过。

2004年，麦格劳法官败选，新当选的法官正是梅西捐助次数最多的个人：本杰明。

“本杰明”更经不起搜索。在美国最高法院的网页上，更多的故事浮出水面。4月8日，有网民贴出了其涉嫌偏袒梅西公司的案件链接：在一场涉及5000万美元的索赔官司中，他在陪审团投出了关键的一票力挺梅西公司。但原告不服，继续上诉到最高法院，2009年6月8日，美国最高法院认定本杰明接受过梅西公司直接和间接逾250万美元的竞选捐款，存在“重大偏袒”的可能，判决其回避。

至此，一幅充满数据细节的图像已经清晰地出现，灾难发生的原因

不言自明。梅西公司利用申诉制度对抗政府的监管，一再拖延被关闭的命运；同时通过政治捐款在权力机关安插、扶持代言人，以争取最大的生存空间。

但这一切，似乎都是在合法的框架内发生的，梅西公司钻的是制度的漏洞。

在数据挖掘的过程当中，网民还共享了其他一些数据和信息：梅西公司下属的另外一家煤矿——鲁比煤矿（Ruby Energy Mine），同样岌岌可危。2009年，鲁比煤矿受到573次警告、1668408美元罚款，总数甚至比出事的上大支矿还多，唯一不同的是，鲁比煤矿受到的严重警告仅15次，较上大支矿少33次，这可能是它还没有“引爆”的原因。

4月8日，矿难发生后的第3天，真相似乎已经全然水落石出。网络舆论、新闻媒体开始转向，开始了煤矿管理制度层面的反思和批判。如何完善制度、避免同样的悲剧在其他煤矿重演，一时间成为新闻舆论讨论的重点。

但没几天，《查尔斯顿邮报》（*Charleston Gazette*）的新闻记者沃德（Ken Ward Jr.）又投下了一枚重磅炸弹。

原本趋于平静的矿难追责突然再次沸腾，而且“U”形大转弯、掉头回到起点，政府的责任再次受到拷问。

# 默认公开推定：和矿难赛跑的原则

每一个公民，如果你想要保护我们国家优秀、硕果累累的遗产，保护自由人民的权益，都应该把政治当做自己的副业。[05]

——德怀特·艾森豪威尔（1890-1969），第34任美国总统

沃德是一位专门从事煤矿报道的知名记者，多次获得新闻大奖。对于这次矿难，他指出了一个细节：矿山安全健康局的每次检查，不仅有以数据格式保存的流水监管记录，还有一份来自检察官个人的详尽文字报告。政府公开了数据格式的检查结果，但却隐藏了这份文字报告。

4月13日，迫于各方的问责压力，矿山安全健康局最终公布了这份报告。沃德随即在后续的报道中引用，大量新的、惊人的事实呈现在公众的眼前。

矿难发生的3个月前，1月7日，美国矿山安全健康局的检察官对出事的上大支矿进行了例行检查，发现了通风设施存在严重的问题，他在检查报告中记录了有关人员对安全隐患令人吃惊的漠视。

这位检察官写道：

“煤矿存在致命的问题，但每个人都在简单漠视：我把一些矿工拉到一边，私底下问他们，你知道通风存在问题吗？他们告诉我：‘我知道’。我还质问过管理人员，但回答是——没有问题，不用担心。

我又问监工摩尔（Terry Moore），他无可奈何地说：我也知道这个问题，我调到这个岗位3周之前这个问题就出现了。我也向矿长反映了，他说：没问题，不用担心。”

这位检察官把他发现的问题和这些对话都记录在他的正式报告里，交差复命了。当天，上大支矿被处以近年来金额最大的一次罚款，梅西公司立即对此进行了申诉。

沃德的报道见报后，举国哗然。

这份详尽的文字报告为什么不予公布？民众在质问。

大部分人都认为，如果政府及时公布这份文件，像沃德这样的专业记者就会对梅西公司漠视安全的行为进行报道，从而引起广泛的社会关注，激起整个社会的“集体行动”，灾难就可能得以避免。

但拖了3个月，政府无所作为，最后一声巨响，等来了矿难。

美国科学家协会的保密专家阿夫古德（Steven Aftergood）对此提出了批评：这是生死攸关的信息，政府应该立即发布。按照现行的法律，政府无法立即关闭煤矿，但政府如果即时发布这些信息，社会就可能通过“集体行动”来阻止悲剧的发生。可惜的是，政府截留了这些信息，堵死了后路。

非政府组织“政府开放联盟”的主任麦得摩（Patrice McDermott）说：煤矿公司为了利益，永远不会告诉矿工真相，政府的义务就是对此进行制衡，制衡的方法，就是信息公开。[06]

一片质疑声中，也有人冷静地提出，矿山安全健康局截留这份文件其实也有“法”可依，不能承担全部的责任。追本溯源，视线不可避免地又投向了《信息自由法》。

1966年，约翰逊总统签署了《信息自由法案》。1977年，美国煤矿管理进行了改革，新的《煤矿法》和《信息自由法案》遥相呼应，规定

煤矿安全的检查结果和详尽的检查报告必须立即公开。1993年10月，克林顿主政时期，司法部部长雷诺（Janet Reno）又发布了著名的雷诺备忘案：

类似刑法中对犯罪嫌疑人的“无罪推定”原则，政府的所有信息适用于“默认公开推定”：即“可以公开”是对所有文件的一个默认假定；任何文件一经产生，就将被默认为“可以公开”，如果有人认为该文件不能公开，必须陈述提供特别的理由。

多年来，美国矿山安全健康局一直践行雷诺备忘案，坚持在第一时间发布、公开煤矿的安全检查信息，像沃德这样的行业记者总是可以很快获得大量来自政府部门的第一手资料，撰写自己的煤矿安全报道。

“9·11”发生以后，布什政府以国家安全的名义，开始对信息进行管制。2001年12月，司法部发布了新的埃斯克劳特备忘案（Ashcroft Memo），否定了雷诺的“默认公开推定”原则。埃斯克劳特备忘案陈述说：

“我们将致力于维护我们的基本价值观，保卫我们国家的安全、商业信息和个人隐私……当你认真考虑了信息公开的要求、并决定不公开的时候，你完全可以相信，司法部将会捍卫你的决定，除非你的决定明显缺乏法律依据或侵害了其他部门的权益。”

在这样的支持和鼓励下，不难想象，联邦各部门以“国家安全、商业秘密和个人隐私”为名，开始先后截留公共信息。一些原来可以立即发布的信息变成了提交申请才能公开的信息，一些原本提交申请之后可以公开的信息，却成为拒绝公开的保密信息。

煤矿安全检查的文字报告就是其中一例。2004年6月起，矿山安全

健康局正式宣布，只公布检查的数据流水记录，对于煤矿安全检查的文字报告，不予公开，即使有公民或组织依据《信息自由法》提交查询的申请，也一律拒绝。

颇具黑色讽刺意味的是，对矿山安全健康局截留检查报告的做法，煤矿公司的老板也怨声载道。有官员回忆说，新做法实施的当月，电话都被打爆了。煤老板们纷纷抱怨说，你只给一个罚款结果，不公开整个调查过程报告，我无法判断该不该对此提起申诉。

奥巴马上任以后，强调建设开放透明的政府，特别对开放公共数据提出了一系列新的要求。2009年3月，在他的推动下，美国司法部推翻了埃斯克劳特备忘案。

但百密一疏，煤矿安全检查的报告也不属于数据，便继续堂而皇之在既定的轨道上运行。

历史还原了真相。奥巴马虽然有决心进一步改革积弊、建立更透明的政府，也做出了一系列信息公开、数据开放的举措和姿态。但冰冻三尺，非一日之寒，行政机关官僚主义的惯性阻碍了改革的脚步。最终，他的改革还是没有跑赢西弗州的这一场矿难。

4月15日，奥巴马在白宫发表了专门讲话，对西弗州的矿难作了基本结论。他认为，矿难的原因主要有两个，首先是梅西公司管理不善；二是复杂冗长的申诉制度纵容了生产隐患长期存在、政府没能及时关闭上大支矿。奥巴马督促国会立即研究新的法规，堵塞煤矿公司利用申诉制度来拖延处罚、继续生产的漏洞。在谈到政府监管时，他要求矿山安全健康局积极检视当前的工作程序和方法，以确保煤矿安全。




珍尼特·雷诺

美国第78任司法部长，也是首任女司法部长。

她为美国政府首次明确了信息发布中的“默认公开推定”原则。她在雷诺备忘案中阐述：“公众对于政府工作的知情权是美国民主的基石。”雷诺在任8年期间，还大力推动“信息和数据”在司法工作中的应用，ComSta就是在她任期中发明、推广的。图为其1996年登上《时代》杂志的封面。

几经峰回路转，至此尘埃落定：29名矿工失去了生命，煤矿公司有待追究法律责任，立法机关要查缺堵漏，但作为主要监管部门，美国矿山安全健康局却并没有人为此丢官受责。

这个结果公平吗？为什么呢？

前文提到过著名的经济学家斯蒂格勒，除了“政府俘虏理论”，他也

为信息经济学做出过巨大的贡献，他有个“信息不对称理论”，可以很好地解释其中的原因。

该理论指出，掌握信息多的人，在社会竞争中处于有利的地位，而信息贫乏的人，则处于不利的地位。前者可以通过向后者有偿转让信息来获取利益；无偿出让，则收获“善意”；只有信息自由流动，各方掌握相同的信息之后，彼此才能建立完全的信任。

考察回放整个事件的细节，不难发现，在矿工、煤矿公司、政府监管部门三方的互动博弈中，关注的焦点就是煤矿是否安全的信息。

政府作为全体公民的代表，可以凭借合法的手段获取几乎全部的信息，是社会的信息枢纽。

煤矿公司处于政府和矿工之间，为了追求利润，它有强烈的动机制造“信息不对称”，向矿工隐瞒真相；但当它面对政府的监管时，它也害怕“信息不对称”，在矿山安全健康局截留检查报告时，喊出了：只给罚款结果，不公布整个调查过程报告，这不公平！

矿工群体则处于斯蒂格勒描述的不利地位：信息链的底层，只能凭借其他方开放的信息来保护自己的安全。

美国煤矿法明确规定，所有的信息公开都是为了保护矿工的安全。矿山安全健康局每次履行监管职能之后，都依法公布检查结果的数据记录。这种免费、及时、主动的信息公开，不仅改变了矿工群体“信息不对称”的劣势地位，如斯蒂格勒所说，还向全社会传达了一种善意、营造了一种信任。

这种善意和信任，最终使矿山安全健康局作为政府监管部门能在舆

论追责中化解压力、渡过危机。

归根结底，如果把信息公开仅仅解读为公众对政府一种单方面的监督，是狭隘的、片面的，信息公开也是政府的一种自我保护。在西弗州矿难追责的过程中，可以看到，政府通过信息公开有效地获得了社会的信任。这种信息公开，不仅对矿工、煤矿公司是一种保护，对政府和整个社会都是一种保护。

既然信息公开能使博弈的三方甚至整个社会都从中受益，为什么还有阻力呢？

这种阻力，又如何才能克服呢？

## 唯一的道路：民主时时都要“争”

在所有的政府组织形式中，民主是最需要能量和想象力的一种形式，也是最需要公民个人公共精神的一种形式。[07]

——乔治·马歇尔（1880-1959），美国军事家、政治家，五星上将

透视矿难前后发生的点点滴滴，我们可以清楚地看到，阻力和问题还是来自于政府。

奥巴马上台以后，通过推行透明行政、开放数据等一系列措施加大了政府公开的力度，但矿山安全健康局并没有顺势主动开放煤矿安全检查的文字报告，还是固守陈规，在原有的惯性轨道上继续滑行，间接导致了矿难的发生。

艾伦·史密斯（Ellen Smith），是美国《煤矿安全与健康简报》的主编，多年来，她一直致力于煤矿安全的分析和报道、强调信息公开在预防安全事故中的作用。[08]她对矿区安全事故专业、详尽的分析，提供了一般新闻报道没有的“观点和细节”，甚至成为矿工和矿区对簿公堂的呈堂证据，受到了法官的欢迎和肯定。

作为矿区安全事故方面的权威，西弗州的矿难发生之后，新闻界对史密斯进行了采访，通过她的讲述，一幅更庞大、充满更多细节的美国煤矿信息开放历史出现在读者的面前。

上文提到，2004年6月，矿山安全健康局宣布将封锁安全检查文字报告。之后，史密斯撰写了大量的文章对布什政府的倒行逆施进行批评，这份文字报告一直是她呼吁公开的重点。她在2004年7月份的文章

《被架空的信息自由：公众有权知道决策背后的原因》[09]中，发出了凌厉的质问：

“这是美国吗？这届政府怎么能够如此践踏民主的传统？矿山安全健康局怎么能说变就变，在没有征询任何公众意见的情况下，一举抛弃自己27年透明开放的传统？”

不仅通过自己的简报，史密斯还会同美国联合矿工工会在《华盛顿邮报》等媒体上公开质疑和呼吁。

2006年1月2日，西弗州发生了一宗死亡12人的矿难，史密斯在她的分析简报中，再次提出煤矿安全检查文字报告的重要性。这一次，她的简报引起了国会的关注。

2006年1月，资深众议员韦克斯曼（Henry A. Waxman）引述史密斯的分析，向劳工部部长赵小兰（Elaine L. Chao）提出质询，要求允许公民和社会组织查阅煤矿安全检查的文字报告。

韦克斯曼在信中说：

“正如一份出版物所指出的：在没有任何公众参与讨论的情况下，矿山安全健康局秘密改变了过去长期定期发布煤矿安全检查文字报告的做法，背后的原因，我们不得而知。但可以肯定的是，这份文件并不涉及国家安全和个人隐私。而且，这个政策为什么改变？其相关文件也从来没有公布过。”

他要求矿山安全健康局检讨一切和信息自由相关的政策和措施。

韦克斯曼的要求，很快得到了另外几名众议员的附议。1月份，先后有多名议员向劳工部发出了公开信，一致表明支持公开煤矿安全检查

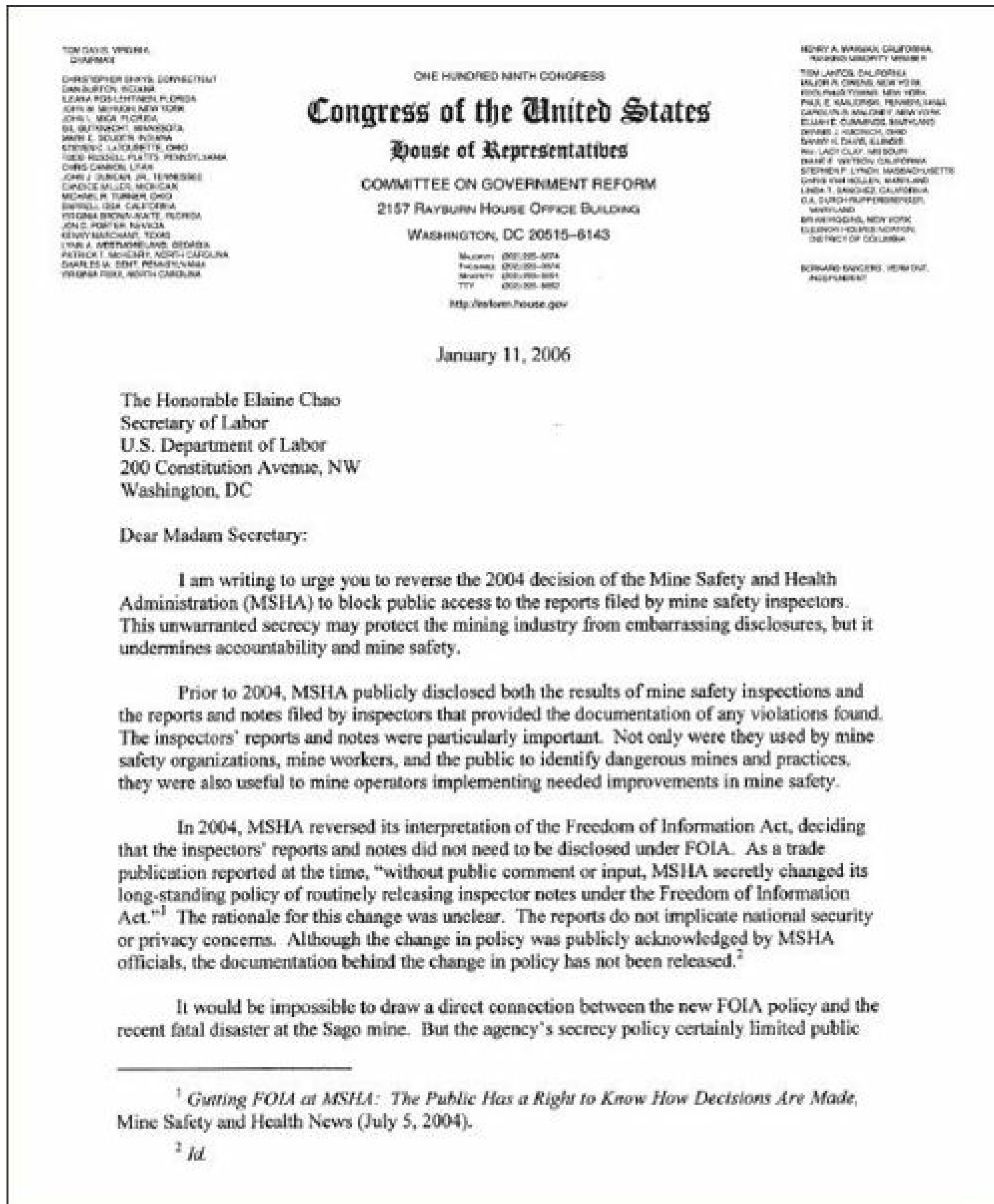
文字报告的态度。

在国会议员的直接施压之下，2006年1月30日，矿山安全健康局终于宣布，公民和组织可以根据《信息自由法》查阅煤矿安全检查的文字报告。

根据这个规定，任何组织和个人都可以通过电子邮件、信件及上门等形式，向矿山安全健康局提交查阅该文字报告的申请，但走完这个程序，需要2周到3周的时间。

史密斯在第一时间发布了这则消息，她称之为公众获得的“一个巨大的胜利”。但她也指出，这个胜利还远远不够。她认为，这份报告堪称矿工的“生命线”，它的发布速度，是在和“矿难”赛跑。它不仅仅应该公开，而且应该即时发布：当检察官完成了检查，当矿工和记者回到家的时候，他们能在网上点击数据流水记录的同时，可以看到检察官对于矿区安全的检查过程和文字描述。

史密斯深信，只有这样发布信息，才能激起矿工、记者、工会和整个社会及时的集体行动，规避可能发生的悲剧。



ONE HUNDRED NINTH CONGRESS

**Congress of the United States**

**House of Representatives**

COMMITTEE ON GOVERNMENT REFORM

2157 RAYBURN HOUSE OFFICE BUILDING

WASHINGTON, DC 20515–6143

http://reform.house.gov

January 11, 2006

The Honorable Elaine Chao
Secretary of Labor
U.S. Department of Labor
200 Constitution Avenue, NW
Washington, DC

Dear Madam Secretary:

I am writing to urge you to reverse the 2004 decision of the Mine Safety and Health Administration (MSHA) to block public access to the reports filed by mine safety inspectors. This unwarranted secrecy may protect the mining industry from embarrassing disclosures, but it undermines accountability and mine safety.

Prior to 2004, MSHA publicly disclosed both the results of mine safety inspections and the reports and notes filed by inspectors that provided the documentation of any violations found. The inspectors' reports and notes were particularly important. Not only were they used by mine safety organizations, mine workers, and the public to identify dangerous mines and practices, they were also useful to mine operators implementing needed improvements in mine safety.

In 2004, MSHA reversed its interpretation of the Freedom of Information Act, deciding that the inspectors' reports and notes did not need to be disclosed under FOIA. As a trade publication reported at the time, "without public comment or input, MSHA secretly changed its long-standing policy of routinely releasing inspector notes under the Freedom of Information Act."[1] The rationale for this change was unclear. The reports do not implicate national security or privacy concerns. Although the change in policy was publicly acknowledged by MSHA officials, the documentation behind the change in policy has not been released.[2]

It would be impossible to draw a direct connection between the new FOIA policy and the recent fatal disaster at the Sago mine. But the agency's secrecy policy certainly limited public

[1] *Gutting FOIA at MSHA: The Public Has a Right to Know How Decisions Are Made*, Mine Safety and Health News (July 5, 2004).

[2] *Id.*

众议员韦克斯曼给劳工部部长赵小兰的公开信：要求公开煤矿安全检查的文字报告。

此后，史密斯继续以笔为旗，四处鼓呼。等到2009年1月，奥巴马就任第一天的开放姿态曾经点燃过她的希望。但史密斯很快就发现，矿山安全健康局还是同样的机械和麻木。2009年5月，她又在简报中批评说：

“虽然我们的新总统对信息公开有伟大的计划和设想，但迄今为止，矿山安全健康局在这方面还没有丝毫的动静和改变。”

西弗州4月5日的矿难发生之后，史密斯扼腕叹息：你见过谁踢自己的屁股？即使我相信奥巴马政府有开放政府的诚意，但政府的信息公开永远不可能自动发生，我们要“争”。

对于政府固有的机械、冷漠和消极作为，史密斯在接受采访时举出了更多的例子。

矿山安全健康局的现任局长梅恩（Joe Main），曾经是美国联合矿工工会的领袖，他和史密斯两人为了煤矿安全，多年来一直并肩战斗、共同奔走。2000年，梅恩被任命为美国矿山安全健康局局长。司法部推行埃斯克劳特备忘案之后，史密斯满怀希望向梅恩提出公布煤矿检查报告的要求，但他的回答却给她一泼迎面冷水：“这涉及隐私问题，我以前也没想到”。

史密斯回忆说：“当年他在工会工作的时候，可不是这么说的。”她后来感叹说：人性的善变，远远超出我们的想象。我们要记住，无论何党当政，何人主政，政府总是有设法掩盖事实的倾向，信息公开不会自动发生，公民要参与、要抗争。

现代政治学的常识告诉我们，史密斯是对的。



艾伦·史密斯

《煤矿安全与健康简报》的主编。从1987年以来，她致力于煤矿安全的分析和报道，已经获得了31个大小新闻奖项。2007年，因为她与矿山安全健康局关于信息公开的抗争，美国记者协会授予她新闻简报公共服务奖（Sigma Delta Chi Awards）。

即使在美国这种民主国家，还是要争。

美国，确实在践行民主制度。但民主，只是政府的一种组织形式、一个框架。每一条法律、每一项政策、每一个决定，还要在这个框架之内讨论、争辩、博弈，才能达成一个共识、输出一个结果，各方利益才能平衡。这个过程，说白了，还是“争”。如果有一方不“争”，那他的利益自然得不到充分的照顾和保证。

此外，民主框架的本身，也是一个“争”的结果。就算已经建立了，它还需要公民不断地去捍卫、去维护、去“争”，防止它在现实中脱轨走样。正如美国作家麦克利什指出的：

“民主，永远不是一件已经完成了的事情。民主是一个过程，需要

一个国家永不停懈的努力。”

也就是说，民主不是结果、而是参与的过程，这个过程，时时都要“争”。说白了，世界是现实的，其实没有一件事情可以坐等、可以一劳永逸。

对于一个非民主状态的政府和社会，当然更要“争”。在专制的框架之内，权力被少数人垄断。任何权力，从少数人垄断演变到多数人的共享，都是一个“争”的结果。从英国中产阶级自发的自由大宪章运动，到引发美国立国的抗税风波，再到近年来台湾地区的民主化，古今中外，概莫能外。

回顾人类社会从古到今所有政府的历史，无论民主也好、专制也罢，所有的政府有一点是共通的：因为缺乏竞争，官僚体制与生俱来都有一种僵化保守的本性，政府机关也往往因此固守不前。但这种僵化保守也不是一成不变的，很多时候，官僚们也需要制造新的口号、接受新的概念来“收买”民意、顺承时代。说到底，一个社会，普罗大众有没有公民精神、是否勇于承担公共责任、争与不争，才是最重要的。

注释

01 原歌词为法文，本书采用了萧三在1923年的译词。1926年，国民革命军第三军曾将此歌译成军歌，其相应译文为：“从来没有什么救世主，不是神仙也不是皇帝。更不是那些英雄豪杰，全靠自己救自己！”
02 To the Families of Upper Big Branch Mine: Our Hearts Ache Alongside You, Obama, April 25, 2010
03 英语原文为：“Every citizen should be a soldier. This was the case with the Greeks and Romans, and must be that of every free state.”—Thomas Jefferson
04 该赞美诗为《放低你的百合》（*Lay the Lily Low*）。
05 英语原文为：“Politics ought to be the part-time profession of every citizen who would protect the rights and privileges of free people and who would preserve what is good and fruitful in our national heritage.”—Dwight David Eisenhower
06 Was Lack Of Government Transparency A Factor In Mine Deaths? Huffington Post, 04-14-2-10
07 英语原文为：“Democracy is the most demanding of all forms of government in terms of the

energy, imagination, and public spirit required of the individual."—George C. Marshall

08 史密斯是《煤矿安全与健康简报》（*Mine Safety and Health News*）的主要经营者，该简报是一份不从属于任何煤矿公司、游说组织及政治机构的独立煤矿新闻月刊。

09 Gutting FOIA: The Public Has A Right to Know How Decisions Are Made, *Mine Safety and Health News*, Editorial, July, 2004

# 外　篇　天下趋势
# 大数据社会未来之世界走向

你可以阻挡一支入侵的军队，你无法阻挡一种思想。[01]

——维克多·雨果（1802-1885），法国思想家，1877年

预测未来最好的方法，就是去创造未来。[02]

——彼得·德鲁克（1909-2005），现代管理学之父

注释

01 “An invasion of armies can be resisted; an invasion of ideas cannot be resisted.”—Victor Hugo, 1877

02 “The best way to predict the future is to create it.”—Peter F. Drucker

# 第十一章　大趋势

原始数据，现在就要！[01]

——蒂姆·伯纳斯－李

万维网之父，2009年2月3日在TED大会上的演讲

# 数据权：大不列颠的硕果

新的“数据权”最令人激动。这将确保人民有权向政府索取各式各样的数据，用于社会创新或者商业创新。通过这些措施，我们可以创建一个最开放、最负责和最透明的政府。让我这样做一个总结：你会有足够多的信息来了解政府是如何运行的、如何花钱的，以及我们工作的效果。使用这些数据、开发这些数据，让我们负起责来，一起努力，创建一个现代民主的典范。[02]

——戴维·卡梅伦，英国首相，2010年11月19日

另一片大陆，欧洲。

和美国相比，开放数据的运动在这里，虽然时间上略有滞后，却大有后来居上之势。

2006年3月，英国著名的《卫报》刊登了一位专栏作家的文章：《把皇冠上的明珠还给我们》。作者把公共数据比喻成“皇冠上的明珠”，他在文章中说：

“我们通过交税支持政府收集公共数据，但当我们想使用这些数据的时候，却还要为此付钱。有一个政府——美国，他们向全社会免费发布数据。各种各样基于地图的信息服务、公共数据的整合，已经在美国兴起，这不是偶然的。美国政府的态度是：用纳税人的钱收集的数据应该免费提供给纳税人使用。更多的研究证明，英国对于数据的封闭态度已经导致我们失去了一些商业机会，一些科研项目，例如气候的变化，也因此受阻。”[03]

这篇文章拉开了英国数据开放运动的序幕，此后，《卫报》在该报的技术版面长年开设了一个“自由数据”（Free Our Data）的专栏，在全社会推广开放数据的概念。

但文章作者推崇的“美国模式”，却在英国引起极大的争议。

争议的焦点在于“免不免费”。英国的一些政治领袖、技术精英都不认可美国的免费模式。他们认为，数据的价值，关键在于质量，而“免费”几乎就是低质量的代名词；只有收费，才可能有高质量的保障。他们主张在政府和社会之间成立一个公共数据公司（Public Data Corporation），专门负责数据的加工和整理，确保质量，再以收费的形式向社会发布。

虽然在关键问题上存在分歧，但数据开放运动随后在英国的发展轨迹，却和美国大致相似，个别细节，甚至“雷人”地相同。

在英国的这场运动当中，也有一位像昆德拉一样的数据英雄，他就是被誉为“世界万维网之父”的蒂姆·伯纳斯－李（Tim Berners-Lee）。



蒂姆·伯纳斯－李

英国人，万维网之父，美国国家科学院院士，全世界知名的数据开放倡导人，2004年曾被评选为最伟大的英国人。近年来，他定居美国波士顿，在麻省理工学院任教。（图片来源：维基百科）

伯纳斯－李早年是程序员。1989年，他开发设计了全世界第一个网站、发明了万维网，为全世界网络资源的组织和访问制定了统一的格式和标准。但伯纳斯－李的伟大绝不仅仅在于技术上的发明和创造，他还是一名社会活动家、新知识的推广者和普及者。他在发明万维网之后，放弃了该项技术的专利，这极大地推动了全世界互联网空间的发展。1994年10月，伯纳斯－李还在美国的麻省理工学院成立了万维网联盟（W3C），主导、推动各种网络标准的建立，并研究、探索下一代网络的具体组织形式。

在对下一代互联网的研究过程中，伯纳斯－李深深地意识到数据对于未来社会的重要性，他说下一代互联网本质上就是“数据网”（Web of Data），他在各种场合大力推广“数据网”的新概念。

2009年2月，伯纳斯－李受邀在TED大会[04]上发表演讲。他的演讲阐述了下一代互联网和开放数据的关系。他说，你想象不出政府会找出多少个借口来拒绝开放数据；不仅仅在美国，全世界都是如此，不仅仅是政府，企业也是这样。他向听众建议说，我们要练习如何索要数据——这很重要！

程序员出身的伯纳斯－李，一直以“内向”、“羞怯”闻名于英国的新闻界，但这一天，他在演讲台上带领全场的听众一句一句地喊口号：

“原始！”

“数据！”

“现在！”

“原始数据，现在就要！”

此后不久，伯纳斯－李和英国的首相戈登·布朗一同出席一次颁奖典礼。布朗向他请教，英国政府应该如何利用互联网，伯纳斯－李立即回答说：把政府的数据推上互联网。他的直言建议，引起了首相的重视。

2009年6月，美国的Data.Gov刚刚上线，伯纳斯－李被英国政府任命为内阁顾问、主管数据开放工作的项目主任。

万维网

（World Wide Web，WWW）

万维网的概念，不同于互联网。我们通过互联网访问的，是一个又一个通过网络相连的资源，这些资源通过一个“统一资源标识”（URI）相互区分，又通过超文本文档（Hypertext）的形式互相链接。当用户点击这些链接，这些内容就通过一个标准的传输协议（HTTP）传送给我们。这一套在互联网上组织资源、获取信息的方法和体系，称为万维网。可以说，互联网一般指的是硬件意义上的网络，万维网才是网络的灵魂。

和昆德拉一样，伯纳斯－李也很快遭遇到了政府部门的铜墙铁壁。几乎每一个部门对开放数据都疑虑重重，拒绝公开的理由也各不相同。伯纳斯－李借鉴了美国阳光基金会的做法，他在英国也举行一次公共数据的应用程序开发大赛，这个竞赛成了转折点。2009年夏天，他带着普通人开发的若干程序，在白金汉宫给全体内阁部长做展示，以此彰显全民数据共享的效果。

他最终赢得了足够的支持。

2010年1月，英国政府的数据开放网站正式出台：Data.Gov.uk。除去地理信息之外，该网站公布了3000多项民生数据。而这个时候，美国的Data.Gov虽然已经经营了半年多，还仅仅只有1000多项民生数据。

英国数据开放网站的重装出台，也引起了美国的舆论哗然，专业人士对两个网站进行了一番里里外外的评头论足，美国的报纸最后提出了问题：为什么别人一出手，数据量就是我们的3倍？

这引起了英美两国在数据开放方面更多的对比和竞争。伯纳斯－李后来将这种竞争称为两国之间“美丽的竞赛”（Beautiful Race）。

2010年5月，戴维·卡梅伦领导的保守党在英国的大选中取得了胜利，但因为没有在议院取得半数以上的绝对多数，保守党被迫和其他的政党组成了联合政府。卡梅伦出任首相之后，他领导的联合政府不仅全面沿袭了上届政府的数据开放运动，还继续深化，提出了“数据权”（Right to Data）的概念。卡梅伦指出，“数据权”是信息时代每一个公民都拥有的一项基本权利，并承诺要在全社会普及“数据权”。不久后，英国女王在议会发表演讲，也强调要全面保障公众的“数据权”。

此后，数据开放俨然成为英国新政府的一个前进方向。

和奥巴马一样，卡梅伦首先瞄准的是公共支出的数据。他上任一个月，就向全社会开放了英国政府2005年以来公共开支的全部原始数据。

2011年4月，英国劳工关系部、商业部又宣布了一个旨在落实、推动全民数据权的新项目：“我的数据”（MyData）。该项目认为：“你的数据，你可以做主！”即使是由商业机构出资收集的数据，但如果记录的是你的信息，你就应该有权查看、使用。在两个部门的主导下，已经有谷歌、巴克莱信用卡、汇丰银行、Groupe Aeroplan、Home Retail Group等十多家不同行业的大公司加入了这个项目，承诺将对社会开放公司收集的与客户相关的数据。

随着越来越多的商业组织开放自己的数据，五花八门的新型应用层出不穷。

欧洲的一家高科技公司Vision Smarts开发一款新的手机程序：Pic2shop。在你购物时，你只需要将智能手机的照相镜头对准商品的条形码，该程序将立刻告诉你，这件商品在其他一些商家的价格以及购买者的评价。在这个分析和对比的基础上，消费者可以马上做出是否购买的决定。




图为Pic2shop的手机界面：显示同一个DVD在不同商家的价格。

可以想象，Pic2shop将给消费者带来多大的经济收益和便利！

又例如，在英国，有几百种手机套餐可供消费者选择。到底哪款套餐最适合呢？这取决于消费者上网、发短信、听音乐、下载、传送图片以及网内、网外通话等各种活动的长短和多少。英国一些手机运营商已经开放了用户消费的明细数据，而第三方立即设计出专业的算法，迅速为用户在几百个套餐当中找到最贴身、最经济的方案。

类似的新应用难以计数，而且每天都在增多。

英国的商业部认为，通过深化“我的数据”项目，将鼓励正面的市场竞争，消除个别商家利用客户“信息不对称”、打“擦边球”的赢利行为。社会资源的配置将更加精细、更加优化，社会运行的总成本将会降低。同时，新的数据开发工作将创造新的就业机会，可谓既开源又节流，全社会都受益。

伯纳斯－李在谈到商业领域的数据开放时曾表示，一个新的、巨大的市场已经轮廓初现，这个市场的规模和潜力，超出了我们的想象。他鼓励英国继续深化数据开放的运动，他认为，这种开放，也是一种竞争，因为开放得早、开放得多将会成为一个国家的先发优势。凭借这种优势，英国的商业机构就能捷足先登，开发出新的应用，这些应用，不仅能惠及本国人民，一旦时机成熟，还可以销售给其他国家。

“我的数据”这个项目，把数据开放的理念从公共领域推进到了商业领域，仅凭这一大步，新闻舆论都认为，英国的数据开放运动已经把美国甩在了后面。这个巨大的进步，在美国也获得了肯定和赞誉。

为了保持在这场运动中的领导者地位，奥巴马也很快推出了新的措施。

# 大合流：国际开放联盟

每个国家都在追求一条自己的道路，这条道路，根植于这个国家人民的文化当中。但是，经验告诉我们，历史的发展最终是站在自由这一边的。开放的经济、开放的社会和开放的政府，是人类社会之所以能够进步最深厚、最强大的基础。[05]

——奥巴马，在联合国大会上的演讲，2010年9月23日

就好像一艘旗舰，当Data.Gov驶出港湾的时候，它缓缓前行，面对一片质疑、反对的声音，作为舵手，联邦政府的首席信息官昆德拉并没有减缓马力，面朝大海、不断前行，空间越来越开阔、越来越高远。终于，批评的声音在慢慢消失，取而代之的是波涛般的掌声和共鸣。

随着Data.Gov上开放的数据越来越多，美国联邦政府的做法，开始在州政府和地方政府的层面得到追随和响应。2009年5月起，先后有加州、纽约州、密歇根州等31个州和芝加哥、亚特兰大、费城等13个大城市先后推出了各自的Data.Gov数据开放门户网站。

一个舰队的阵容已经形成。

这支舰队，也获得了来自国际社区的阵阵喝彩，除英国之外，加拿大、新西兰、德国、法国都开始纷纷效仿，推出了自己国家的公共数据开放网站。

作为一名雄心勃勃的政治家，奥巴马知道，他可以适时再迈进一步，把他“建立前所未有开放政府”的理念推进到国际社会，争取国际空间的支持。

2010年9月23日，奥巴马作为美国总统，在联合国大会上发表了演讲。

联合国大会是一年一度的联合国议事会议，由各成员国的首脑或高级代表参加，就重要的国际问题发表意见和观点。

奥巴马在这次演讲中说，我们面对的，不是一个普通的时代。虽然各国的人民有不同的文化、面临不同的挑战，发展路径也各不相同，但归根结底，自由才是历史发展的最终趋势，而开放的政府、开放的经济、开放的社会正是人类之所以能够不断向前发展、获取更大自由的根本动力。

他号召各个国家跟上全球技术创新的脚步，拿出开放政府的具体措施：

"一个开放的社会可以支持开放的政府，但是却不能替代开放的政府。自由选择领导人、自己决定自己的命运，这是人类最基本的一项权利。现在，我们要看到：并不是美国的要求和推广导致了民主在世界各地的成功，民主在一个社会的到来，是因为这个社会的每一个公民对社会管理开始发言并要求拥有一份决定权。

……

我们还可以看到，在世界各地，创新正在为政府的开放和问责带来新的机遇。这时候，我们必须有所作为。当我们明年在联合国大会重聚的时候，应该对一些问题，例如如何提高透明度、打击腐败、鼓励公民参与、利用新的技术巩固国家的根基，带来具体的承诺和措施，拥抱那些正在点亮我们这个时代的新思想。"[06]

奥巴马的演讲，无疑是美国联邦政府新的号角。之后，Data.Gov明显加快了国际化的步伐。

2010年11月15日，美国商务部、内务部共同组织了第一次开放政府数据的国际会议，来自英国、新西兰、澳大利亚、巴西等十多个国家的100多名代表参加了这次会议。会议的口号是：我们将共创历史（We Will Make History Together）。

昆德拉和伯纳斯－李自然是这次会议的主角，在这次会议上，英美两国还正式形成了数据开放的伙伴关系。英国宣布，它们将尽快筹办第一届开放政府数据研讨会。[07]

一周之后，来自全世界30多个国家的100多名代表齐聚伦敦，召开了第一次开放政府数据的研讨会。

2011年的10月，第二届研讨会移师波兰。来自全世界41个国家一共250多名程序员、社会活动家、政府官员、公益领袖和记者参加了这次会议，在热烈的讨论中，酝酿了一个新的战略转变。

会议指出，全世界已经有50多个大大小小的数据开放网站，仅仅再增加网站的数量，已经意义不大，要引导开放数据运动在世界范围内的深化，应该着眼于建立全世界统一的数据开放平台和开放标准。

美国迅速对此作出了反应。

2011年12月，美国联邦政府宣布，将和印度政府共同合作，把现有的Data.Gov改造成开源平台，在2012年开放全部的平台代码。源代码发布之后，全世界任何国家都可以免费引进、使用及修改美国的数据开放平台。印度将率先移植Data.Gov，作为其中央政府的数据开放平台。

这个新的决定，再一次让美国站在了全世界数据开放和创新的潮头浪尖。

这一次，这一创举的背后，除了昆德拉之外，还有另外一股力量。

美国国务院也在为奥巴马的开放战略积极拓展国际空间。

2011年7月，美国国务卿希拉里和巴西外交部长帕特里奥塔（Antonio Patriota）在华盛顿达成协议，计划以奥巴马倡导的开放理念为基础，共同发起一个新的国际组织。不久后，英国、挪威、墨西哥、印度尼西亚、菲律宾、南非陆续加盟，发起国增加到8个。

2011年9月20日，这8个国家在纽约集会，宣布成立“开放政府联盟”（OGP），并发布了《开放政府宣言》[08]。8大发起国在宣言中誓言将用自身的行动来推动世界各国政府的开放，并许下了4大承诺。第一大承诺便是：向本国社会开放更多的信息。

宣言书说：

“政府代表人民收集并保存了各种各样的信息。人民有权利获取关于政府活动的各种信息。我们承诺：用可以重复使用的格式，及时主动地向社会开放高质量的信息，包括原始的数据。”

申请加入开放政府联盟的门槛——参与国必须具备以下**4**个条件

| 1．财政透明 | 2．信息自由 | 3．财产公开 | 4．公民参与 |
|---|---|---|---|
| •能及时公开政府预算和支出的信息和文件 | •有专门的法律保证公民获取政府信息和数据的权利 | •有专门的制度规定如何公开高级政府官员的收入和财产 | •允许公民参与公共政策的制定、保护民权 |

说明：可以看出，这4个基本条件都是奥巴马担任参议员和总统之后的主要政治主张。

奥巴马在这次会议上发表了开幕辞和闭幕辞。

第二天，9月21日，他又在联合国大会发表了一年一度的演讲。他这次演讲的主题是“和平”，但他还是利用这个机会，向各国政府的代表解释了发起“开放政府联盟”的原因。他说，要通过这个组织，和其他的国家联手，共同利用开放社会和开放经济的力量，提高社会对政府的问责度、激发社会和公民的潜力。

几个月后，开放政府联盟又陆续收到了加拿大、意大利、希腊、韩国等42个国家/地区的加盟申请，其会员迅速增加到50个。其中，有31个国家/地区都建立了公共数据的开放网站（统计日期为2012年4月25日）。

在这31个国家/地区当中，还不乏发展中国家。

例如，本书一开始就提到的非洲穷国肯尼亚。2010年8月，肯尼亚通过了新的宪法，其宪法第35条规定：

“每一个公民都有权获得政府拥有的信息。……每一个公民都有权修改、删除（政府保存的）不真实、有误导倾向的错误信息。”

公民的信息权写进宪法，信息的开放自然是水到渠成。2011年7月，肯尼亚总统齐贝吉（Mwai Kibaki）宣布推出公共数据开放网站：opendata.go.ke。随着这个网站的建立，肯尼亚成为非洲大陆第一个开放数据的发展中国家。截至2011年底，肯尼亚的网站上共开放了390组数据。

中国的香港地区也融入了这股大潮。2011年3月，香港推出了公共数据开放网站：Data.One，目前该网站只开放地理和交通两大类数据，

已有近百组数据可供下载。

全世界目前已经正式建立数据开放门户网站的国家和地区

| | |
|---|---|
| 美洲 | 秘鲁、加拿大、美国、乌拉圭、智利 |
| 亚洲 | 阿联酋、巴林、东帝汶、韩国、沙特阿拉伯、新加坡、中国香港 |
| 欧洲 | 爱尔兰、爱沙尼亚、奥地利、比利时、丹麦、德国、法国、芬兰、荷兰、摩尔多瓦、挪威、西班牙、希腊、意大利、英国 |
| 非洲 | 肯尼亚、摩洛哥 |
| 大洋洲 | 澳大利亚、新西兰 |

# 云计算：新的航向

就像公用电话网一样，计算的能力，有一天会被组织起来，成为一种公共资源和公共事业，这种公共资源和事业，会成为一个新的、重要的产业。[09]

——约翰·麦卡锡（1927-2011），美国计算机科学家、图灵奖获得者，1963年

除了奥巴马的理念创新，昆德拉还在技术上不断突破，精心打造Data.Gov这艘巨舰。

2011年5月，联邦政府宣布，将推出第二代Data.Gov。新的方向，是为Data.Gov建立一个以云计算为基础的平台。

“云计算”（Cloud Computing），是一个较为抽象的概念。其来源和演变，凝聚着众多科学家的智慧和创新。

“云”的概念来源于电话通讯行业。20世纪，电话已经普及成一种公共事业，通过公共电话网，两个点之间可以实现通讯。但如果想在两个用户之间建立一条专用的、私密的通讯渠道，则必须架设新的物理专线，这相当于另起炉灶，非常昂贵。1990年代，一种被称为“虚拟专用网络”（VPN）[10]的技术出现了，它可以通过公用网络随时为两个用户建立专线联系，这是一个革命性的突破，大大节省了通讯的开支。为了形容这种可以为个人提供专用资源，并可以招之即来、挥之即去的网络服务，“云”的概念产生了。

“计算”，指的是计算机的计算能力，其大小快慢，取决于计算机内

硬件的配置。人类发明计算机不久，计算机科学家就开始憧憬，把计算能力集中起来共享。1963年，人工智能的另一位先驱、斯坦福大学的约翰·麦卡锡教授就预见说：“计算的能力，有一天会被组织起来，成为一种公共资源和公共事业。”1983年，太阳微系统（Sun Microsystems）的首席研究员约翰·盖奇进一步明确说：这种组织，就是网络。[11]也就是说，把计算的能力放在互联网上，而不是你桌面的个人电脑上。所有的硬件计算能力、存储能力、软件执行能力，全部都由网络提供：网络就是你的计算力，网络就是你的电脑。

这种通过网络将计算能力组织起来的做法，可以实现经济学意义上的“规模化”和“专业化”，意味着巨大的利润空间。

1997年，南加州大学的一位印度裔教授切诺柏（Ramnath Chellappa）将“云”和“计算”组合成一个新的单词，正式提出了“云计算”的概念，他认为：“从此以后，计算的边界将由经济的规模效应决定，而不仅仅取决于技术层面的限制。”

这些天才的思想自然引起了企业界前赴后继的尝试。

1999年，美国著名的客户关系管理软件提供商Salesforce提出了利用云计算提供软件服务的概念（即软件即服务）。之后，戴尔、亚马逊、IBM、谷歌、微软等公司纷纷加入，投入巨资，开发自己的云服务。其中，亚马逊公司拔得头筹，2006年，该公司成功地推出第一款提供硬件服务的云产品，即“弹性云计算”（EC2）。

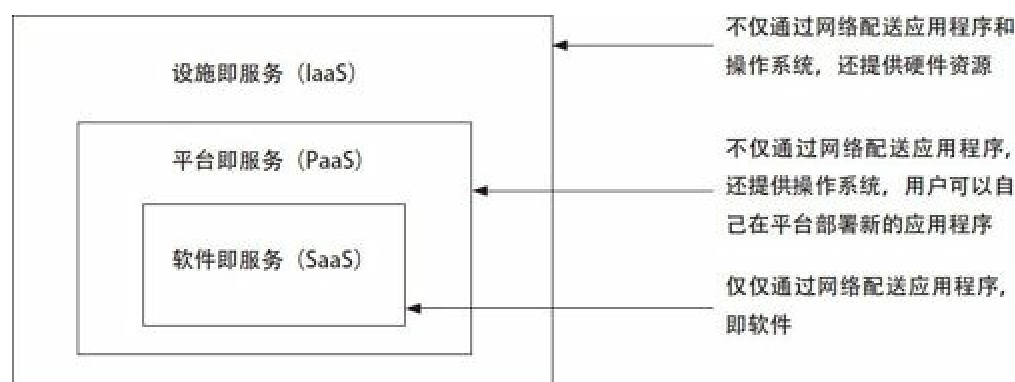
2010年前后，云计算已经形成了一个从应用软件、操作系统到硬件的一个完整产业链，开始得到大规模的商业应用。

云的运营商把计算能力当做一种资源，集中在一起，然后再通过网

络，配送给有需要的客户。客户需要的计算资源多，服务商就送得多；客户需求一下降，配送就可以立刻下调。客户购买这种服务之后，就不再需要再购置额外的软件和硬件。这对小公司而言，意味着不用投入大量的资金购买服务器和软件，通过租用，立刻就可以享受到以前只有大公司才能配置的软硬件能力。

也就是说，就像自来水管道供水、电力网输电一样，云计算把“计算”从有形的产品变成了无形的服务。计算能力成了一种可以传送的服务，这是继互联网的出现之后，信息技术领域最重要的一个创新和变革。

云计算服务的三种模式




说明：三种模式分别对应应用软件、操作系统和硬件，服务范围层层扩大。SaaS、PaaS和IaaS分别代表Software as a Service，Platform as a Service和Infrastructure as a Service。

1993年，互联网刚刚出现不久，克林顿总统高瞻远瞩、提出了后来举世闻名的信息高速公路计划，在其任期内为美国搭建了一个覆盖全国的光纤通信网络，这为美国全面进入信息时代奠定了基础，当然也拉动了经济发展、提升了美国的综合国力。其后世界各国纷纷仿效，在全球

掀起了一股“信息高速公路”的热潮。

潮起潮落，十几年过去，云计算又带来了新的机遇，将引发信息产业的重新布局。雄心勃勃、对创新和科技情有独钟的奥巴马当然不愿错失这个历史机遇。2010年，奥巴马在工业界和学术界聘请了71名专家，成立了“云”委会，帮助联邦政府普及“云”知识、制定“云”政策、推动“云”部署。

当然，联邦政府云计算背后的主要推手，还是首席信息官昆德拉。

对于这个新的概念，大部分联邦政府的官员都曾经“云山雾罩”、人云亦云，并不清楚什么是真正的“云”。

作为联邦政府的首席信息官，昆德拉曾经在大小场合解释什么是“云”以及“云”的优势，他对“云”的描述，非常生动、形象：

“曾经，每一个家庭、农庄、村落、城市都必须有自己的水井。今天，你仅仅打开水龙头，干净的水就通过公共供水管道输送给我们。云计算，也是这个道理，就像我们厨房里的水一样，可以根据我们的需要，随时打开或者关上。在自来水供应公司，有一群专业人员负责水的质量、安全以及24小时不间断供应。当你关上水龙头，你不仅节约了水，你也不用为你没用的水付费。”

2010年12月，在昆德拉的主导下，联邦政府宣布了“云优先”（Cloud First）政策，规定所有新建的政府信息系统，必须优先考虑云平台。2011年2月，白宫正式发布了《联邦政府云战略》[12]，该《战略》要求，各个国家部委必须确定3个可以推向云平台的系统，并在年内完成至少一个。同时，联邦政府2094所数据中心，通过云计算，在2011年合并了137所，未来5年将再精简800所。

作为开放政府的旗舰，Data.Gov已经成为奥巴马政府的一张名片，自然要走在创新的前沿。2011年5月，Data.Gov宣布开始迈向云平台的改造，对联邦政府而言，这意味Data.Gov将变成一种服务，数据、软件、服务器都将保存在私营公司的平台上，用户可以在平台上开发、部署、运行自己的应用程序，服务的收费，将取决于存储量、计算量、访问量等等指标。

2011年8月，昆德拉辞去了联邦政府首席信息官的职务，他接受了哈佛大学的邀请，前往伯克曼互联网和社会研究中心（Berkman Center for Internet and Society）从事技术和政治的研究工作。美国报纸评论说，无论是他领导的数据开放运动，还是他为联邦政府制定的云战略，都把美国政府推向了创新的潮头，领跑了整个世界。

美国白宫发布的一个经典案例:为什么需要“云”？

2009年6月，为鼓励消费，联邦政府推出旧车换新车的现金补贴项目（Car Allowance and Rebate System），该项目申请和审批的过程都在网上完成。交通部开发了软件、装备了服务器，并预计全美将有25万人上网申请。

不料仅仅一个月，申请人数就高达69万，系统不堪重负、一再瘫痪，导致了诸多社会抱怨。白宫后来总结说，如果是云平台，不仅可以立即增加系统的资源和容量，任务完成之后，资源还可以立即释放，另作他用。

奥巴马正是希望借助“云”的伸缩性，实现公共资源的“按需配置”，这不仅可以节约资金，还可以提高公共服务的质量，同时获得更多的决策弹性，大胆创新。

云计算的出现，把数据存储和数据分析变成了一个可以更加方便获得的网络服务，毫无疑问，这是一个重大的变革。随着它的普及，全世界政府、企业和个人使用、消费信息技术的模式，正在改写。

但大数据时代，还在催生更多的变革。

一个划时代的、更大的巨变，正在慢慢向我们靠近。

## 再造互联网：从网页相连到数据相“联”

语义网就是数据网，从某种程度上来说，就像一个全球性的数据库。……语义网不仅仅是把数据放上网，它还要在数据之间创建联接，数据一旦联接，计算机和人都可以对数据进行探索：通过一个数据发现另外一些相关的数据。[13]

——蒂姆·伯纳斯－李，1998年、2006年

大数据时代正在催生的最大技术变革，是重新构造互联网。

对下一代互联网的研究，伯纳斯－李领导的麻省理工学院万维网联盟（W3C）是全世界的领跑者。1998年，该组织就开始提出下一代互联网的定义和构建方法。

伯纳斯－李将下一代互联网称为“语义网”（Semantic Web），他继而解释说，语义网就是“数据网”（Web of Data）。

所谓“语义”，是指遵循一个统一的标准，给每一片信息赋予一个计算机都能理解的“意义”，这个标准的“意义”，用术语说，则是“元数据”，也可以形象地理解为给信息贴上各种标准化的“标签”。

如前文所述，现在互联网上的资源组织形式，是万维网（World Wide Web）。在万维网上，网页是信息资源的基本组织单位，各个网页都有一个地址，即“统一资源标识”（URI），它们通过开发者定义的链接连接起来，用户可以从一个网页跳跃切换到另一个网页，即网上冲浪。

在语义网上，数据将像网页一样，成为组织资源的单位。一个数

据，可以像万维网上的网页一样获得一个地址（即统一资源标识），同时，还有统一的语义对它进行描述。

这样，语义网上的数据，就不再是一个死的数字，而是一个活的“细胞”，它可以被定位，还拥有和其他数据语义一致的标签，这意味着它可以和其他数据相联。之所以称为相“联”，而不是相“连”，是因为，它们并不是像网页一样通过一个链接简单连在一起，而是通过数据之间内在的关系挂起钩来、“联”在一起。这种关系，不是随便定义的，而是基于数据的含义和属性产生的。

元数据（Metadata）

元数据是大数据时代的一个重要概念，它是指描述、解释数据属性的数据（data about data），是为支持一致性的数据描述所定义的统一准则。

这种相联，就像两个数据库通过“主键”（Primary Key）相联起来一样，不同的是，这里的“主键”，是一个数据的元数据。

对用户而言，这意味着可以从一片数据自由地跳转到其他数据——数据冲浪。这个跳转，依靠的不是人为的链接，而是本质关系上的联结。

这将是一次划时代的革命。

1990年代，互联网进入大众的生活之后，人们开始在网上大建网站，网站的目的，是信息的传播和分享，这个时代，很多人称之为“Web 1.0”。随着技术的进步，21世纪初，出现了推特、脸谱[14]等社交

网站，社交网站把基于互联网的交流、协同和互动演绎得淋漓尽致，被称为“Web 2.0”。

但在大数据时代，信息共享、交流互动已经不再是最迫切的需求，数据的分析和整合，才是最大的挑战，而这恰恰是万维网的短板。

可以说，万维网是一个网页的集合体。这些网页，即使内容相似、主题相同，但由于分属不同的网站、存储在不同的服务器上，他们之间就没有链接和联系，这意味着如果用户想找到它们、分析它们，就必须借助搜索引擎等工具。

这也是谷歌之类的搜索公司为什么能成为商业巨头的根本原因：在现在的互联网上，各种信息之间是孤立的，我们要对它们进行整合、分析，就必须依靠人工的外力，目前最好的工具就是搜索引擎。

但搜索引擎并不完美。通过特定的计算机算法，搜索引擎按“关键字”的相关度对网页进行过滤排序，然后给用户返回一大批相关的网页，这些信息，良莠不齐，还需要进一步的人工分辨。

未来的语义网，却不是一个众多网页的链接体，而是一个全球性数据库。在这个数据库中，各种相关数据通过“元数据”互相联结，计算机将根据元数据，自动为我们搜寻、检索和集成网上的信息，不再需要搜索引擎。

例如，在万维网上，如果你想阅读韩寒的作品，你可能要借助搜索引擎，输入“韩寒的作品”，但搜索引擎返回的，是一板一眼包含“韩寒的作品”这5个字的所有网页，其中，有韩寒创作的作品，也有别人对韩寒作品的评论等等相关的信息。但在语义网上，韩寒创作的作品会有一个语义标签，别人的评论会有另外一个标签，通过标签，韩寒所有的作

品将自动相联，你找到了一个，就能发现一片。

又比如，在语义网上，你可以跨网站整合数据：你收到银行寄来的消费记录，但却记不清其中一笔10元钱是不是自己花的，你可以立即将你的银行记录和你的Outlook日程表联结起来，通过逐条查对你的行程、活动和消费明细，这样你很快就能想起，那天花钱的时候，你正在和家人开车去公园的路上，你买了几瓶水。跨网站整合数据，还意味着你可以将微博上的一条信息群发给你开心网上定义的朋友圈。

不难想象，在新一代的“语义网”里，谷歌的搜索服务，将不再有立足之地，各种社交网站的功能和格局也将改写，语义网的智能平台将触发、呼唤一些崭新的服务。由于这些变化，也有人将语义网称为“Web 3.0”。

从**Web 1.0**到**Web 3.0**的演变




说明：Web 1.0到Web 3.0的提法在业界有人支持、有人反对，并不统一。

但语义网的建设却非一日之功、一人之力。

首要原因是有大量关于元数据的标准需要制定、统一。

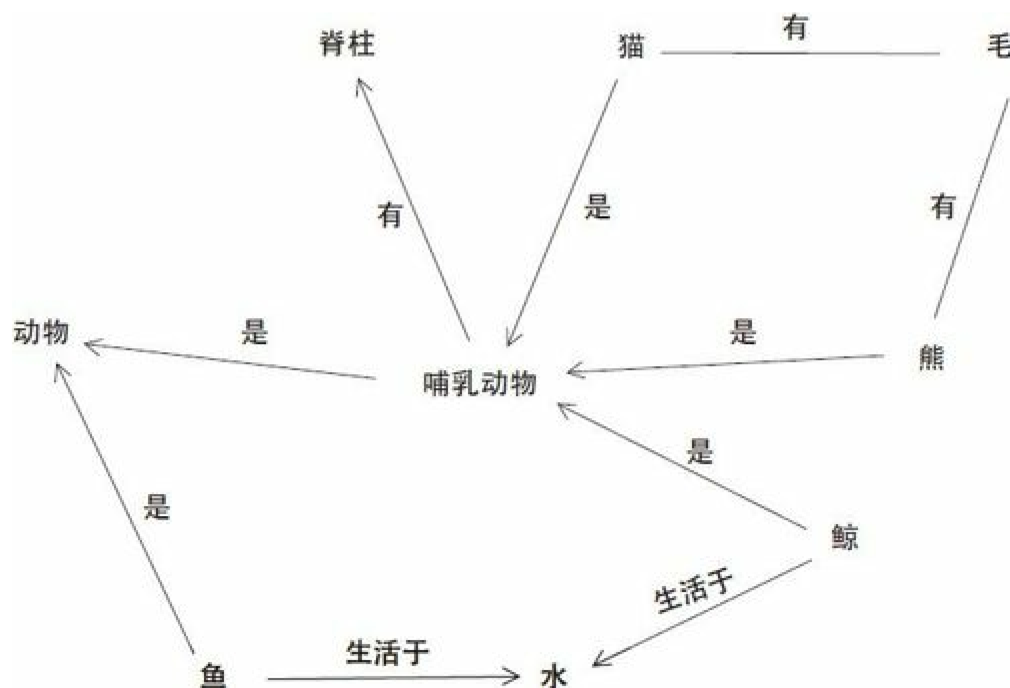
语义网的建设，元数据是关键。只有通过这些计算机能理解的语义

标记，每一个片断的信息才可以和其他的信息自动发生联系，揉捏整合，形成联结的关系。作为一种“标记语言”，元数据必须是标准的、一致的，又是灵活的、可扩展的，适用于万事万物。

为了保持“标记语言”的一致性，计算机科学家正在对全世界万事万物的基本特点及其关系进行规范和定义，这个过程，叫做创建本体（Ontology）。所谓本体，就是世界上所有的常见事物。创建本体，也可以理解成对人类社会每一个实体进行概念化、标准化的定义过程。

可以想象，这是一个如何浩大的工程。

简单的本体示例：关于动物的概念及其相互关系所构成的语义网络




说明：通过以上这几个本体定义，计算机一碰到“熊”，就知道它是一种哺乳动物、身上有毛，还能推断出它有脊椎、不在水中生活。（图片来源：维基百科）

在各种各样本体的基础之上，为数据加上语义一致的元数据标签，这是语义网最基础也是最繁琐的工作之一。以Data.Gov为例，昆德拉在创立之初，就鼓励政府各个机构按照语义网的标准发布数据，目前，该网站上有400多组按语义网的标准发布的数据，但仅仅这几百组数据，其定义的元数据就达到64亿之多。

语义网蓄势未发，除了基础工作庞大繁杂之外，还有一个经济学上的原因，可以用网络的外部性来解释。

何为网络外部性？

网络外部性

（Network Externality）

网络外部性是理解现代经济的一个重要概念。它是指某件产品对单个消费者的使用价值，取决于这件产品总用户的多少。用户的人数越多、形成一个越大的网络，每个用户从产品中得到的效用就越高。

电话、即时通讯软件、社交网站等等都是网络外部性的典型例子。

让我们穿越到1876年，电话刚刚被美国人贝尔发明的时候。

假设安装一部电话非常昂贵。当你计划安装时，你会审视你周围的朋友，如果他们都不买、也不装，那你的电话就无人可打、毫无作用，只有用的人越多、你能联系的人也就越多，电话对你的价值才越大。语义网也一样，你投入资源、人力，把你的网页、数据重新定义了，但还

不够，你的数据必须和其他数据相联，你的工作才能产生价值。也就是说，你投资的效果，不是立竿见影的，而是最终取决于其他人是否投资、和你相联。只有全球性的数数相联，语义网才能最终形成、发挥最大的效用。正因如此，目前，商业领域对语义网的投资很有限，语义网的建设主要靠政府和公益组织在推动。

在2009年的TED大会上，伯纳斯－李的演讲题目是《下一代网络》，他在向观众描述了未来语义网的强大之后，向听众呼吁：

“现在，我们正处在这样一个阶段：如果你认为这是一个伟大的想法，那你就要去做。这件事，每个人的回报将取决于其他人是否行动。我想，今天我们在座的很多人，他们做事，并不是因为这件事很快就有投资回报，他们去做，是因为他们明白：每个人都去做了，就能成就一件好事。”[15]

换句话说，语义网的建设，也存在着“集体行动的困境”。

今天，当我们面对互联网时，好像感觉不到任何的变化。放眼望去，大地似乎平坦依旧、江河仍然向东奔流，但地表之下，正在平静地发生剧烈的变化。每一天，世界各地都在定义新的本体、增加新的互联数据。这些变化，也像地基，地基打得越深越牢，楼才能盖得越高越大。

但这些本体，主要还是英文本体，中文世界对本体的研究和贡献极为有限。

终有一天，全球将数数相联。

就像浏览网页已经成为我们的生活方式一样，数据在网上的自动整

合和跳转，将会成为我们新的生活方式。通过这种新的方式，互联网将向我们“推送信息”，而不再是我们在网上“搜索”信息。

那将是一个更加辉煌的大数据时代。大数据之“大”，将不仅仅意味着数据之多，还意味着，每一个数据都能在互联网上获得生命、产生智能、散发活力和光彩。

## 注释

01 英语原文为：“Raw Data Now!”—On the next Web, TED 2009 Conference, Tim Berners-Lee, Feb 3rd, 2009

02 英语原文为：“The most exciting is a new right to data, which will let people request streams of government information and use it for social or commercial purposes? Take all this together and we really can make this one of the most open, accountable and transparent governments there is. Let me end by saying this. You are going to have so much information about what we do, how much of your money we spend doing it, and what the outcome is．So use it, exploit it, hold us to account. Together we can set a great example of what a modern democracy ought to look like.”—David Cameron, November 19th 2010

03 Give us back our crown jewels, *The Guardian*, Charles Arthur and Michael Cross, March 8, 2006

04 TED是Technology、Entertainment、Design（技术、娱乐、设计）3 个英文单词的缩写。TED大会是全世界著名的思想领袖会议，该大会以美国加州为基地，每年将众多的技术、设计、文学、音乐等领域的杰出人物集聚在一起，对科技和文化发展中的重大话题进行交流、思考和探索。

05 英语原文为：“Each country will pursue a path rooted in the culture of its own people．Yet experience shows us that history is on the side of liberty; that the strongest foundation for human progress lies in open economies, open societies, and open governments.”—Remarks to the United Nations General Assembly, Obama, September 23, 2010

06 美国总统奥巴马在联合国大会上的演讲，2010年9月23日。

07 Open Government Data Camp: 2010

08 Open Government Declaration, September 2011

09 英语原文为：“Computing may someday be organized as a public utility just as the telephone system is a public utility.…The computer utility could become the basis of a new and important industry.”—John McCarthy, speaking at the MIT Centennial in 1961

10 互联网上流行的“翻墙”技术，用的其实就是虚拟专用网络的技术。

11 英语原文为：“The network is the computer.”—John Gage, Vice President of Sun Microsystems, 1983

12 Federal Cloud Computing Strategy, Vivek Kundra, U.S. Chief Information Officer, February 8, 2011

13 英语原文为：“The Semantic Web is a web of data, in some ways like a global database. …The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other related data.”—Tim Berners-Lee, 1998 & 2006

14 推特即Twitter，相当于中国的微博；脸谱即Facebook，是美国最流行的社交网站，相当于中国的开心网和人人网。

15 On the next Web, TED 2009 Conference, Tim Berners-Lee, Feb 3rd, 2009

# 第十二章　大挑战

在全球经济的很多领域，大数据在以很多的方式创造价值。事实上，研究表明：随着消费者、公司、各个经济领域不断挖掘大数据的潜力，我们正处在一个巨大浪潮的尖峰，这个浪潮，就是大数据驱动的创新、生产率提高、经济增长以及新的竞争形式和新的价值的产生。[01]

——《大数据：下一个创新、竞争和生产率的前沿》

麦肯锡全球研究所，2011年5月

上一章谈到，在大数据的时代，数据将成为一种权利、数据开放将成为一种潮流，计算正在从一种必须购买的产品转变为一种可以随时租用的服务，互联网将最终演变为智能的数据网。

在这些正在发生和演变的趋势当中，大数据的时代，也蕴藏着对个人、企业或组织甚至全社会的巨大挑战。

以本书的主人翁之一奥巴马为例，他的第一个任期即将结束，按照美国的民主制度，他正面临着能否再次当选的挑战，这次选举，也可以说是一次关于数据的竞争。

# 逐鹿政坛：得数据者得天下

政治技术决定政治成败。[02]

——莫顿·布莱克韦尔，美国政治活动家

2011年4月4日，奥巴马以电子邮件和视频的方式，正式宣布自己将参加2012年的总统竞选、争取连任。

4年多以前，2007年2月，他在广场上的人潮中宣布了自己将参选总统的决定。作为代表伊利诺伊州的联邦参议员，他选择了位于该州首府的议会广场作为集会地点，这里曾经是林肯总统发表演讲的地方。奥巴马在凛冽的寒风中，面对沸腾的人群和高声喝彩，发表了近一个小时的演讲，随后开始了他长达一年多的大选征程。




2007年2月10日，奥巴马在伊利诺伊州的首府斯普林菲尔德

（Springfield）议会大厦的广场上宣布自己将参选总统。（图片来源：Mark Wilson/Getty Images）

但这一次，没有簇拥的集会、鼓动的口号和大幅的标牌，一切都由网上的点击，在瞬间平静地完成。

另外几位热门人选，如前麻省州长罗姆尼（Mitt Romney）、前众议院议长金里奇（Newt Gingrich）也先后用推特（Twitter）和脸谱（Facebook）正式宣布了他们的参选意愿。

而通过集会拉开竞选的序幕，原本是美国上百年的政治传统。

有评论说，这表明，脸谱、推特和视频分享网站的发展已经深刻地改变了美国的政治生态，新的传统正在形成。还有评论指出，这次总统大选，是民主、共和两党都充分认识到社交网站重要性之后的第一次大选，技术对政治的作用将前所未有地凸显。

回顾历史，从上个世纪30年代的广播，到60年代的电视，再到本世纪初的互联网，技术一直在美国的政治生活中扮演重要的角色，甚至决定选举的成败。

互联网出现之后，就交流和互动而言，广播和电视无疑相形见绌。和它们相比，网络能提供一种没有强权的、多中心的、双向的甚至多向的信息交流方式，这一改广播、电视单向交流的不足，使候选人和选民之间的互动方式发生了深刻的改变。

如果说2008年的大选，是一场关于网络技术“用多”还是“用少”的较量，那么2012年的竞争，工具本身已经不再是竞争的优势。因为，近4

年来，民主、共和两党都鼓励本党的政治家把网络作为必争之地，网络的重要作用已经得到了政治家的一致重视。共和党曾经发起“新媒体挑战”（New Media Challenge）的竞赛活动，以网络上粉丝的增长速度为衡量标准，鼓励党员集聚网络人气。比赛进行了一个半月，最后来自路易斯安那州的众议员佛兰芒（John Fleming）夺得了第一名。民主党当然不甘落后，也举办了类似的“在线政治明星”评比活动。

电台总统：罗斯福

除了率先在总统选举和执政中引进民意调查的方法，罗斯福总统还是第一位大规模使用广播电台的候选人和总统。

他对广播电台的利用，集中体现在“炉边谈话”上。

1933年他刚刚上任，正值美国大萧条时期，民间充满了焦虑、挫折和抱怨。

通过广播演讲，罗斯福自信、亲和的声音走进了千家万户，拉近了大众和政府之间的距离，从而在心理上造成了一种休戚与共的神圣感。每当听到炉边谈话，人们就仿佛看见脸上挂满笑容的罗斯福，所以有人说，“华盛顿与他们的距离，就像起居室里的收音机一样近”。甚至有民众将他的照片剪下来，贴在收音机上。

对广播电台的娴熟应用，帮助罗斯福成功当选了4任总统。炉边谈话取得的巨大影响，也成为广播史上的一个传奇。

但广播电台的局限性也是显而易见的：只闻其声，不见其人，某些情况下，声音可能成为一种误导。



（部分资料来源：中文维基百科“炉边谈话”条目）

## 电视总统：肯尼迪

1960年9月，肯尼迪与尼克松进行了美国历史上第一次总统候选人电视辩论。

在电视辩论中，肯尼迪给人的印象是一个准备应对国家重大挑战的领导人，尼克松给选民的印象则是一个试图战胜对手的竞争者。大多数通过收音机收听辩论的民众认为尼克松在辩论中占了上风，但事实恰好相反。所有电视机前的观众都看得出来，是肯尼迪占了上风，他看上去更加轻松，更有自制力。而镜头前的尼克松却显出害怕的样子，脸色阴沉、不停地冒汗。在舞台背景的灯光下，身穿浅灰色西服的尼克松淡化成了一个模糊的人影，而肯尼迪的深色西服却在光线反差中显得十分鲜明。

两人共进行了4次电视辩论。肯尼迪最后以微弱的优势战胜了尼克松，成功当选。

此后，电视成为政治活动中重要的工具。电视弥补了广播“闻其声、不见其人”的巨大不足，但也有其弱点：一个“演”，一个看，信息的传播是单向的，对观众而言，可“望”而不可“及”，没有互动可言。电视直播也常常成为一些政客作秀的工具，美国众议院的一项调查表明，电视直播增加了议员演讲的次数，有些议员甚至故意在镜头前大放厥词、以博出彩。



（部分资料来源：维基百科“肯尼迪”条目）

网络总统：奥巴马

2008年11月，奥巴马借助强大的互联网交流动员能力，战胜了共和党的麦凯恩，当选总统。奥巴马的成功，使他获得了“网络总统”的美誉，也从此开启了互联网在美国政治生活中大规模的应用。

奥巴马在竞选中曾经收集了1300万人的个人信息和邮件地址，他如今拥有2000多万的脸谱粉丝、1000多万的推特粉丝。

麦凯恩在竞选中败北之后，也迅速反省了自己的弱点，开始投入“围脖”。他目前已经拥有170万推特关注者和87万Facebook粉丝，每天都发推特更新。作为一名75岁的政治家，他的转变之快，曾经引起

了美国舆论的惊叹。

奥巴马的成功也不是特例。2010年1月，共和党布朗（Scott Brown）在麻省与民主党候选人克勒（Martha Coakley）争夺联邦参议员的席位。克勒是麻省的司法部长，麻省又是民主党的大本营，布朗和她相比，既无天时，也不占地利。但布朗占了网络上的“人和”：他在Facebook上有23万粉丝，是克勒的10倍。他和当年的奥巴马如出一辙，成功地利用了网络平台进行筹款、招募志愿者，最后当选，成为近40年来在麻省胜出的首位共和党参议员。这一系列网络传奇，令不少学者相信，候选人在互联网上的表现，甚至可以像民意调查一样，成为预测选情的温度计和水晶球。

2011年7月6日，奥巴马在白宫利用推特召开了第一次会议，短短一小时，有6万多名美国网民参与了推特的互动。

2011年7月20日，共和党在推特上举办了美国历史上第一次总统候选人的辩论，共6名候选人参加了推特辩论。在90分钟的辩论中，网民发出了16000多条新的帖子，个别问答被转发了近万次。

2012大选在即，美国的学者、研究机构、新闻媒体无不把候选人的网络人气作为研究选情的一项重要跟踪指标。



说明：2011年7月6日，奥巴马在白宫举行历史上首次推特会议。（图片来源：The Christian Science Monitor）

**2012**年美国总统大选主要候选人的网络影响力排名

| 排名 | 候选人 | 身份 | 推特粉丝数 | 脸谱粉丝数 | 影响指数 | 总粉丝数 |
|---|---|---|---|---|---|---|
| 1 | 奥巴马 Barack Obama | 现任总统 | 13,578,607 | 25,891,809 | 91.2 | 39,470,416 |
| 2 | 罗姆尼 Mitt Romney | 前麻省州长 | 402,897 | 1,554,058 | 85.5 | 1,956,955 |
| 3 | 金里奇 Newt Gingrich | 前众议院议长 | 1,451,187 | 296,253 | 81.5 | 1,747,440 |
| 4 | 保罗 Ron Paul | 现任众议员 | 145,772 | 925,424 | 65.1 | 1,071,196 |
| 5 | 桑托勒姆 Rick Santorum | 前任参议员 | 184,155 | 187,841 | 83.8 | 371,996 |
| 6 | 洪博培 Jon Huntsman | 前犹他州州长、前驻华大使 | 81,782 | 37,393 | 60.8 | 119,175 |

奥巴马已经铁定获得民主党的提名，他在网络的影响力也稳居第一，暂时无人可以超越。其他5位主要的挑战者均来自共和党阵营，其中前众议院议长金里奇是著名的《美国之约》的作者，在1995年曾率领共和党在众议院成功夺回失去了42年的控制权，其后才有本书第6章的掺沙子法案之争。至于前参议员桑托勒姆，他曾提出《国家天气服务职责法案》，反对免费发布天气的数据，这曾经在本书第5章提及。前驻华大使洪博培已经在2012年1月宣布退选。

网络影响指数引用的是美国Klout Score指数，该指数对个人在各大主要社交网站的影响力、活跃程度进行加总，然后给出一个综合评分。以上数据统计截至日期为2012年4月1日。

也就是说，两党的主要竞争者都在大规模地使用网络。竞争的重点已经不是“用不用”的问题，而是“如何用”、“用得好不好、巧不巧”。

互联网在政治竞选中的主要功能有两个，一是交流，候选人通过它和选民展开个性化的实时对话、收集采纳民意，并为自己的政治主张寻找支持；二是动员，通过它发动草根力量，完成召集志愿者、募款、造势等关键的选举活动。

当前，谁能把网络交流和动员这两个功能发挥得更好、运用到极致，其竞争的核心和重点，就是数据。

可以说，这是一场数据竞争。数据收集、分析和整合的能力已经成为技术和网络竞争的关键所在。

2008年，奥巴马通过他的个人竞选网站（Barackobama.com）收集了1300万人的个人信息和邮件地址，这曾经是他在交流规模上无可比拟

的优势。但和2008年相比，如今的脸谱等社交网站已经不仅仅是个交流平台，还是一个记录了8亿人资料的开放“信息港”，第三方可以在上面开发新的应用并与其现有的数据直接相联，这为其他候选人提供了超越奥巴马的可能性。

奥巴马在宣布再参选以后，立即重新改版了他的个人网站。支持者一登录，便被要求提交其社交网站的账号，并询问是否能够读取其在社交网站上的档案信息，甚至索要在用户社交网页上发布消息的授权。这些举措，招招式式都是为了获取更多、更翔实完备的选民数据。

其他候选人当然也不甘落后，几乎每个候选人都先后推出了自己的网站，个别候选人甚至不惜重金，用送出纪念品的形式来鼓励选民把自己的账号和社交网站的账号挂钩相连，其目的都是为了扩展自己的数据容量，扩大未来与选民交流的规模。

与交流规模相比，交流的质量更为重要。大规模发送信息并不难，但只有挠到接收方痒处和痛处的信息才可能引起真正的关注。这就要求发送方必须了解信息受众，而不是简单地“群发”。这个方面，奥巴马已经有教训和体会。2011年7月29日，美国国会在为是否提高国家债务的上限进行激辩。奥巴马一天内连接发出十多个推特（微博），要求他的粉丝采取行动、给议员施压，催促他们通过提高债务上限的提案。但当天晚上，奥巴马的推特就流失了33000多名粉丝。奥巴马的技术团队随后对此进行了调查，发现其中的原因，正是很多人对这个话题不感兴趣，认为“总统在给他们发送信息垃圾”。

奥巴马的团队正在全力提高数据细分的能力。例如，堕胎问题一直是两党辩论的热点，奥巴马如果能在辩论之时，锁定35岁以下的女性进行信息交流，就可能形成稳定的观点联盟，造成声势。又比如，一谈到

养老金问题，50岁以上的老人将成为第一目标群体。

除了性别、年龄之外，选民还可以按地域、种族、教育程度、宗教信仰等其他维度来划分，这些维度，还可以交叉，交叉分析的维度越多，信息受众就能分得越细、瞄得越准，信息交流的效率也就越高。此外，除了数据的细分，还可以通过数据模型预测不同选民的不同需要和行为模式，最终把个性化的短信和视频通过互联网或手机送达特定的群体。

除了交流，成功的动员也离不开有效的数据支持。研究表明，选举动员的重点是争取“墙头草”或“中间派”，最有效的方式是派出志愿者上门游说，这可能将选票拉升几个百分点，在势均力敌的情况下，成为胜出的关键。如果能掌握这类“中间派”人士的基本信息甚至个人喜好，志愿者的登门游说就能有的放矢、事半功倍。所以如何在网上确定这个群体、获取他们的数据，是决定动员效率的一个关键。

自我量化（**Quantified Self**）：数据驱动的个人新生活

本节从奥巴马竞选的角度阐述了大数据时代对个人生活层面的影响。

但数据对个人生活的影响绝不仅仅限于竞选。近年来，数据对个人生活的影响非常广泛，例如，2005年前后，美国开始兴起一股“自我量化”的运动。

自我量化是指利用计算机、智能手机以及各种新的电子便携感应器来记录自己学习、工作、运动、休息、娱乐、饮食、心情等等个体行为的情况，就像我们需要对体重、身高、血压、心跳等物理指标进

行监控一样。自我量化的主张者认为，为了更好地了解我们、提高自己，我们必须要用数据来记录、研究、分析自己的行为。

其理由在于：人的感觉中存在盲点，直觉不可信任，理性思维也有局限，大脑即使有惊人的记忆力，也未必有惊人的信息加工能力。很多时候，我们会高估自己的理性，低估情绪对我们的影响。认识自己虽然很难，但非常重要。基于数据的记录和分析，可以帮助我们走出错觉、认识真正的自己。

2007年，著名技术活动家、《连线》杂志的主编凯文·凯利（Kevin Kelly）和技术专栏作家沃尔夫（Gary Wolf）共同创办了一个网站：quantifiedself.com，该网站已经成为自我量化运动的枢纽。像数据开放运动一样，自我量化的运动现在非常活跃，在美国、欧洲频繁召开各种论坛和会议。

此外，通过不同数据库之间的分析和整合，为毕业于同一个学校、居住在同一个地区甚至拥有同一个爱好或观点的志愿者建立新的网络联系体，让支持者感觉同声同气、彼此相连，也能极大提高动员的效率和士气。

围绕数据细分的这些精细化的应用，可谓多不胜数。这种基于数据的选民细分方法，也被称为“纳米定位”（Nanotargeting）。

当然，奥巴马能否最终当选，还取决于政治、经济、社会等等诸多方面的因素，但如果仅仅从技术层面来考虑，这一仗争的是数据，比的是数据收集、分析以及整合的能力。从这个意义上来说，得数据者得天下。这种基于数据的竞争，当然不仅仅局限在政治选举的领域。其实，早在2008年奥巴马当选总统之前，商业领域的数据竞争现象就引起了学

界的研究和关注。

# 数据竞争：企业赢之道

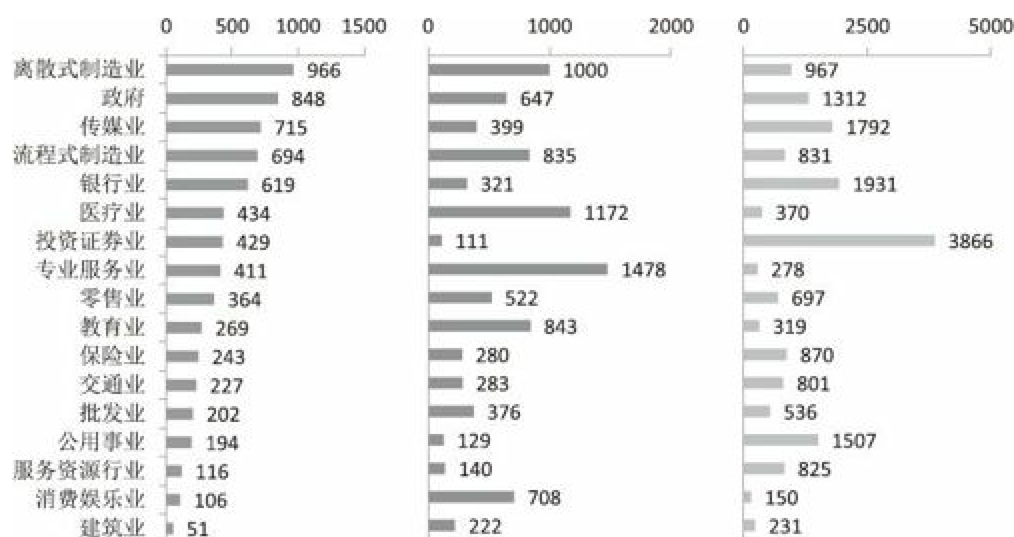
一些公司已经把他们商业活动的每一个环节都建立在数据收集、分析和行动的能力之上了。所有的公司都可以从他们的成功当中学习。[03]

——托马斯·达文波特

美国信息经济领域著名的教授、世界顶尖的商业咨询顾问，2007年

本书第二章提到，虽然美国联邦政府堪称一个数据帝国，但大数据的爆炸，绝不仅仅局限于联邦政府。根据麦肯锡公司的研究和统计，[04] 仅仅美国的制造行业，就拥有差不多两倍于美国政府的数据。此外，美国新闻业、银行业都拥有可以和政府相提并论的海量数据。

作为全世界馆藏量最大的图书馆，美国国会图书馆2009年拥有的电子信息总量为235太字节。从以下的图表可以看出：美国17个行业中，已经有15个行业大公司平均拥有的数据量超过了美国国会图书馆，其中投资证券业、银行业、传媒业位居前三位，都远远超过了美国政府。



说明：投资证券业的数据总量仅居17个行业的第7位，是因为该行业的公司主体较少（该行业拥有的1000人以上的大公司在17个行业中最少，仅111家），如果按公司为单位对比，投资证券业每一个大公司平均拥有的数据量远远超出其他行业，为3866太字节。（图表来源：麦肯锡全球研究所；数据来源：IDC，US Bureau of Labor Statistics，McKinsey Global Institute Analysis）

2010年，全球的企业一年新存储的数据就超过了7000拍字节，全球消费者新存储的数据约为6000拍字节，这相当于十多万个美国国会图书馆的藏书量。每一天，无数的数据在被收集、交换、分析和整合，麦肯锡公司认为，数据已经如一股“洪流”注入了世界经济，成为全球各个经济领域的重要组成部分。麦肯锡公司预计，数据将和企业的固定资产和人力资源一样，成为生产过程当中的基本要素。

这当然是大数据时代的独特现象。和其他的生产要素相比，数据，无疑有其独特的特点。例如，工业生产过程中的原材料，一般都有排他性，但数据很容易实现共享，而且使用的人越多，其价值越大；数据也不像机器、厂房一样会随着使用次数的增多而贬值，相反，重复使用，它可能增值。此外，此数据和彼数据如果能有机地整合到一起，可能就会产生新的信息和知识，而且大幅增值。

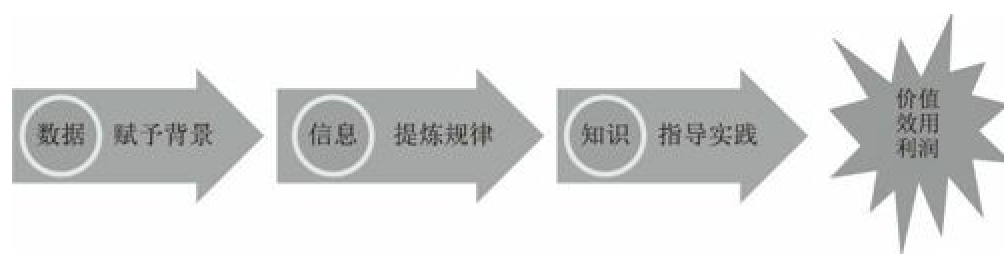
麦肯锡的报告继而指出：

“已经有越来越多令人信服的证据表明：只要实施正确的政策和激励，大数据将成为竞争的关键性基础，并成为下一波生产率提高、创新和为消费者创造价值的支柱。”[05]

这把数据的重要性提到了竞争性要素的高度。如前文所述，信息时

代的竞争，不是劳动生产率的竞争，而是知识生产率的竞争。数据，是信息的载体、是知识的源泉，当然也就可以创造价值和利润，可以预见，基于知识的竞争，将集中表现为基于数据的竞争，这种数据竞争，将成为经济发展的必然。

从数据到利润的跨越和转变




托马斯·达文波特（Thomas H. Davenport）是美国信息经济领域著名的教授，他曾经在埃森哲、麦肯锡等世界知名的咨询公司工作过，对知识经济、数据竞争有深入的研究。他认为：随着全球性竞争的不断深化，企业的地理优势将淡化，各种国家和地区性的保护措施也在逐步取消，一项专利会很快地被模仿、复制、推广，创新将越来越难。但在除去这些要素之外，还有一点可以构成企业竞争的基础，那就是以“低成本、高效率”的方式来开展公司的业务，这种竞争，要求公司制定流线型的商务过程，各个过程之间必须无缝隙、无摩擦地对接，并保证每一个商业决策明智、正确，在竞争的过程中不犯错误。

而要做到“低成本、高效率”的运营以及决策正确，企业必须广泛推行以事实为基础的决策方法、大量使用数据分析来优化企业的各个运营环节，通过基于数据的优化和对接，把业务流程和决策过程当中存在的每一分潜在的价值都“挤”出来，从而节约成本、战胜对手、在市场上幸存。



托马斯·达文波特

哈佛大学博士，现任教于美国巴布森学院。2000年，他被《CIO》杂志评选为“新经济十大杰出人物”；2003年被《咨询》（*Consulting*）杂志评选为全世界“最顶尖的25名咨询师”之一。2007年、2010年他分别出版了两本专著，阐述企业之间的数据竞争现象。

这种竞争，就是一种基于数据的竞争。

达文波特认为，能够始终保证自己以“数据最优”的方式运营的公司将会在竞争中坚持到最后，并不战而胜，因为粗放型经营的公司最终将因为“高成本”自动出局。

这也是孙子兵法中强调的“不可胜在己，可胜在敌”[06]的制胜之道。

达文波特对美国的多个行业进行了考察，发现各个行业都出现了以数据分析为竞争形式的企业。这些企业当中，既有网络时代的新秀，如谷歌，也有经营了百年的品牌，如宝洁。它们都在数据分析的基础之上与其他企业展开竞争，同时又都是行业中的佼佼者。达文波特还相信，这两个特点是相互联系的，也就是说，正是因为这些公司大规模地采用了数据分析的方法，他们才成为行业中的领先者。

达文波特还分析了数据竞争首先在美英等西方国家兴起的主要原因：

“有很多理由促使大部分发达国家的公司争相贯彻数据分析的战略。事实上，这种以数据分析为竞争能力的公司都是各自领域的领袖，他们都把自己的成功归功于对数据分析的娴熟应用。全球性的竞争正在变得更加激烈，这加剧了这种需要。和中国、印度的竞争对手相比，西方国家的公司无法在产品的成本方面获得优势，他们只能在商业过程的优化方面战胜对方。”[07]

各行各业都出现了以数据分析为竞争能力的企业**[08]**

| 行业 | 代表企业 |
|---|---|
| 消费品行业 | 安海斯 - 布希公司（Anheuser-Busch）<br>美国嘉露酒庄（E. & J. Gallo Winery）<br>玛氏（Mars）<br>宝洁（Procter & Gamble） |
| 金融服务业 | 巴克莱银行（Barclays Bank）<br>第一资本金融公司（Capital One）<br>加拿大皇家银行（Royal Bank of Canada）<br>前进保险公司（Progressive Casualty Insurance）<br>Wellpoint 公司 |
| 体育娱乐业 | 奥克兰运动家队（Oakland A's）<br>波士顿红袜队（Boston Red Sox）<br>哈拉斯娱乐公司（Harrah's Entertainment）<br>万豪国际酒店（Marriott International）<br>新英格兰爱国者队（New England Patriots） |
| 工业品行业 | CEMEX 公司<br>约翰·迪尔公司（John Deere & Company） |
| 电信业 | Sprint 公司（美国电信运营商）<br>$O_2$（英国电信运营商）<br>法国电信公司（Bouygues Telecom） |
| 制药业 | 阿斯利康公司（AstraZeneca）<br>苏威制药公司（Solvay）<br>尖端制药公司（Vertex Pharmaceuticals, Inc.） |
| 零售业 | 亚马逊公司（Amazon.com）<br>JCPenny 百货公司<br>乐购（Tesco）<br>沃尔玛（Wal-Mart） |
| 运输业 | 联邦快递公司（FedEx）<br>施奈德公司（Schneider National）<br>联合包裹公司（United Parcel Service） |
| 电子商务行业 | Google 谷歌<br>Netflix 公司<br>Yahoo! 雅虎 |

我们再以美国的零售业为例。

2010年1月，《纽约时报》报道说，由于数据收集意识的增强和新技术的出现，很多零售商拥有的数据量在过去3年里增加了5倍。[09]例如，无线射频技术（RFID）开始在仓储管理中得到大规模的应用，这不仅提高了数据采集的粒度，还增强了数据采集的频度，极大地推动了该行业数据量的增长。这些数据，正在成为零售商最大的财富和资源。

时至今日，数据的收集、分析和应用已经不是该行业局部的现象，而是穿透整个行业、深入骨髓的应用。可以说，零售业的每一个细胞，都在“数据竞争”的空气中呼吸成长。

首当其冲的是顾客分析。每一件流到顾客手中的商品，都会在销售前台产生一条记录，对这些数据进行分析，宏观上可以把握顾客的结构、流量、购买的周期以及不同顾客群的利润贡献率；具体到每一名顾客，可以确定其购买频率，预测其感兴趣的商品、忠诚度和流失的可能性。有了这些由点、线、面齐全的分析，零售商就可以确定最可靠的顾客群体、预测其消费意愿，主动为其提供个性化的销售和促销服务，提高销售额和利润率。

其次是商品分析。核心问题是如何为店面确定最佳的商品组合和库存量，何时导入新商品、淘汰旧商品？既有的商品如何分类上架，什么和什么应该摆在一起或者捆绑销售？节前节后、淡季旺季，甚至一天之内的不同时段每一商品该如何定价？何时降价，以何种形式促销？比如，飓风到来之前，哪些商品要提高仓储量？要提高多少？这些都是要基于数据分析才能回答好的问题。个别的分析不仅需要企业内部的数据，还需要整合来自市场、竞争者以及天气、地理情况等外部资源的数据。

再次是对供应链的分析、对运营效率的分析，这包括选择供应商，优化物流、现金流和配置人力资源等等。

这些分析，是环环相扣、一体化的。举个例子，如果导入不畅销的新商品，滞留货架，超市就可能变成“博物馆”，即使供应环节和运营效率再高，也是一场空忙；又比如，个别商品供不应求，如果库存不足或者供应链滞后，那货架就会空置，错失商机不说，也浪费了空间。

大数据将在**16**个方面改善零售业的运营

| 功能（Function） | 大数据的杠杆（Big Data Lever） |
|---|---|
| 市场营销（Marketing） | 1. 交叉销售（Cross-selling）<br>2. 基于地理位置的市场营销（Location based marketing）<br>3. 店内行为分析（In-store behavior analysis）<br>4. 顾客细分（Customer micro-segmentation）<br>5. 情感态度分析（Sentiment Analysis）<br>6. 多渠道消费体验（Enhancing the multichannel consumer experience） |
| 商品（Merchandising） | 7. 分类优化（Assortment optimization）<br>8. 价格优化（Pricing optimization）<br>9. 上架布局及设计的优化（Placement and design optimization） |
| 运营（Operations） | 10. 绩效透明（Performance transparency）<br>11. 雇员工时优化（Labor inputs optimization） |
| 供应链（Supply chain） | 12. 库存管理（Inventory management）<br>13. 配送和物流优化（Distribution and logistics optimization）<br>14. 供应商谈判（Informing supplier negotiations） |
| 新的商业模式（New business models） | 15. 价格比较服务（Price comparison services）<br>16. 网络营销市场（Web-based markets） |

说明：麦肯锡公司认为大数据将在以上16个领域为零售业带来巨大的机遇，如果公司运用数据得当，零售商的营业利润将提高60％以上。（表格来源：Mckinsey Global Institute Analysis）

林林总总，不一而足。一句话，基于数据的分析和竞争已经是零售业的常态，毫不夸张地说，零售业的竞争一定程度上已经成为一种基于数据的竞争，数据就是企业的财富和金矿，数据分析和挖掘的能力就是企业的核心竞争力。

零售业的这种状态，并不是企业界的特例。无论是横向比较行业的数据总量，还是纵向考察数据分析方法的历史渊源，零售业其实都坐不上数据竞争的“头把交椅”，前排就座的还有金融、保险、投资等领域。但零售行业的这种竞争状态证明，数据的爆炸将影响未来行业的竞争方

式，未来的竞争将更加精细化，这种精细化，就体现在数据的收集和分析上。

这也表明，和互联网时代相比，大数据的时代，不仅意味着更广泛、更深层的开放和共享，还意味着更精准、更高效、更智能的管理革命。在大数据的时代，数据就是直接的财富、就是核心的竞争力，很多行业，都要相继跨入一个数据兴则企业兴、数据强则企业强的竞争时代。

# 下一波浪潮：从大数据到大社会

就像国家本身没有它自己的钱一样，国家也没有它自己的权力，但不幸的是，没有人清楚地明白这个道理。国家所有的权力都是社会授予的，或是以各种理由和借口向社会剥夺的，除了社会，国家权力没有任何其他的来源。[10]

——艾尔伯特·诺克（1870-1945），美国著名作家、社会批评家

美国，作为全世界信息技术最为发达的国家，也是最早迈进信息时代的国家。

信息时代，是人类继农业时代、工业时代之后的一个崭新的文明阶段。在信息时代，信息是比物质和能源更为重要的资源，信息的产生、发布、使用、整合将成为经济活动的枢纽，并对全社会的政治生活、文化生活都产生重要的影响。

1946年，第一台计算机在美国问世，这可以视为人类社会进入信息时代的起点。

1975年，比尔·盖茨创办了微软。次年，史蒂夫·乔布斯成立了苹果电脑公司。个人电脑、商业软件开始逐步走进美国的千家万户。

1993年，克林顿在美国推动信息高速公路计划，千万台终端开始和互联网相连。人人上网，机机相连，信息的交流和共享变得越来越快捷、越来越方便。互联网的普及，是美国社会开始全面迈进信息时代的标志。

但这之后，信息技术的发展并没有停滞。新的技术和观念还在不断

出现，甚至可以用突飞猛进来形容。

近20年来，商务智能、社交网站、无线传感器、云计算、语义网等等新的技术风驰电掣般进入了我们的生活，软件开源、数据开放、普适计算、软件即服务、智慧地球等等新的思想令人应接不暇，如一架高速行驶的列车，我们往往还来不及看清楚窗外的风景，就已经呼啸着越过了下一个站台。

当新技术爆发的能量一次又一次地冲击我们这个社会的时候，不少学者、研究人员惊呼：

我们正在以飞快的速度穿越我们定义的信息时代……列车将驶向何处？

1995年，美国的技术活动家、麻省理工学院的教授尼古拉斯·尼葛洛庞帝（Nicholas Negroponte）出版了《数字化生存》一书，在这本全球性的畅销书里，他率先提出了后信息时代的概念，他认为：

“长期以来，我们都太热衷于讨论从工业时代到信息时代的转变，以致一直没有注意到我们正在进入后信息时代。”[11]

他认为，美国正在进入后信息时代。他总结了后信息时代的几个特点：一是信息变得极端个人化，在后信息时代，每时每刻，都有大量的信息在产生、在流动，但这个时代已经有很强的信息细分能力，虽然有海量的信息在流动，但个人接受到的涓涓细流都是量身订制的；二是计算机能记住个人的细节，机器对人的了解程度将不亚于人对人的了解程度；三是时空障碍将被打破，人们可以在任何地方进行工作和合作。尼葛洛庞帝还认为，个人将在后信息时代获得更大的解放，国家的权力将更加分散、国家之间的界限将更加模糊。

另外一些信息学家和未来学家却认为，后信息时代的最大特点将是创新。信息的保存、加工、发布将不再是经济的重心，基于信息的创新将成为经济的先导。而且由于信息无处不在、人皆可得，后信息时代的创新将不再是少数人的专利，创新将走向民主化，集中表现为终端用户的创新、普通大众的创新。

世界各国的学者都在对这个全新的时代进行展望，美国信息技术学者、创业家维伟克·瓦德瓦则把这个时代称为“新信息时代”：

“以前很难找到的信息和知识可以很快找到，这个历史时期，我们称之为信息时代。但我认为：我们现在已经远远地超出了这个时代，我们正在跨入另一个全新的时代——新信息时代。”[12]

对于这个新的历史时期，虽然还没有统一的名称和定义，但我们可以看到，这之所以能够成为一个崭新的时代，并不仅仅是由于新技术的出现，而是由于新技术的广泛应用和深层渗透，引起了社会形态和社会结构方面的改变，例如大众创新、个人获得了更多的自由、国家权力分散化等等，最终给人类带来更大的解放等等。

我们也不难发现，大数据时代，已经具备了这个新时代在技术层面上几乎全部的特点：

信息高度个性化已经成为现实：商务智能已经可以帮助我们高度细分、定制信息，数据挖掘的技术可以根据用户过去的行为模式预测其信息的需要，为其配送个性化的信息，未来的语义网更是一个高度个性化的智能网络平台！

时空障碍已经打破：因为互联网的普及，在家工作、远程协作已经是平常之事，云计算又将有形的计算转化为无形的服务，计算将像水和

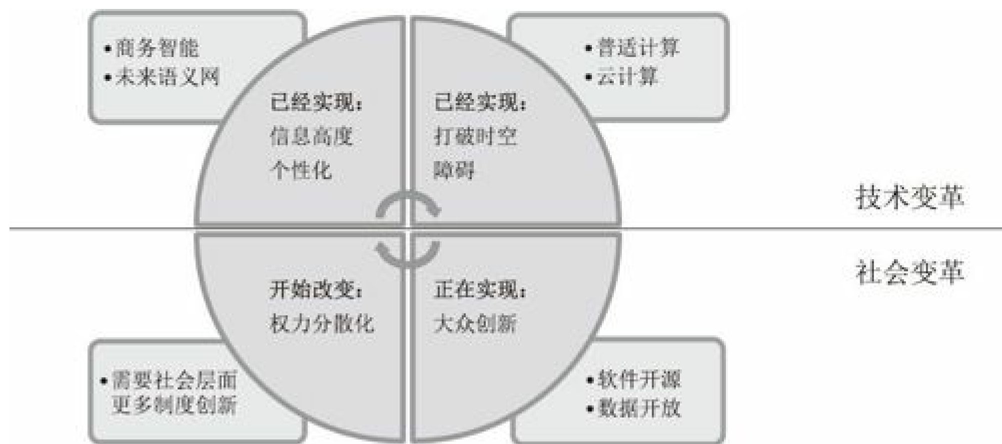
电一样被供应和输送，再随着无线传感器的普及、普适计算的层层深入，“万物皆联网、无处不计算”正在成为现实，时空障碍将被彻底打破！

和技术层面的改变相比，社会层面的变化更加缓慢、微妙，是个渐近的过程。但即便如此，我们也可以清楚地察觉到大数据时代给社会带来的一些改变：软件开源已经逐渐成为一种共识，数据开放也正方兴未艾，这两股潮流正在催生、见证越来越多的大众创新——从1990年代美国街头警察梅普尔的ComStat，到2000年的CitiStat，再到Data.Gov造就的航班延误分析系统、商品召回手机查询系统、数据混搭器等等，无一不是普通人的发明和创造，毫无疑问，在大数据的时代，创新的力量正在向普通大众流动。

即使在权力分散化方面，我们也可以在英美两个国家观察到正在发生改变的端倪。

2010年7月，英国的新首相卡梅伦在深化数据开放运动的同时，推出了一个新的执政纲领。他主张将更多的权力从中央政府下放到地方政府，从地方政府下放到社区，从政治家下放到公民大众，让大众承担更多的责任，以建立一个“更大、更强、更好”的社会。他的新政纲确定了以下5项主要措施：

大数据时代已经具备了下一个新时代的主要特点




1．赋予社区更多权力；

2．鼓励人们在社区活动中成为志愿者，承担更多的责任；

3．从中央政府分权到地方政府；

4．支持民间团体、基金会、慈善组织和社会企业；

5．公开发布政府数据。

卡梅伦认为，他的执政纲领代表一种新的社会文化，公众将拥有更多的自由和权利。例如，邮局、图书馆、博物馆是过去政府承担的典范公共服务，在卡梅伦“大社会”的理念里，民间组织也将拥有自主经营创办这类机构的权利。目前，这种新的社会文化还没有一个固定的名称，卡梅伦解释说：

“你可以叫它自由主义，可以叫它授权激励，叫它自由，叫它责任。我则称它为大社会。”[13]

“大社会”的提法，迅速成了英国社会的舆论热点，甚至被评选为英国2010年的“年度词汇”。

国家和社会之间的权力平衡

国家的权力，集中表现为政府的权力，它其实来自于社会。对于国家、社会以及两者之间权力的分配，美国著名的作家、社会批评家艾尔伯特·诺克曾有一个精彩的论述，他比喻说：就像国家本身没有它自己的钱一样，国家也没有它自己的权力，国家所有的权力都是社会和个体授予的，或是以各种理由和借口向社会剥夺的。社会和个体，才是国家权力唯一的来源。诺克还指出：

“透过公共事务的表面，我们将发现一个基本的事实，这就是：权力在国家和社会之间不断地重新调整和分布。”

每一种社会形态，都对应着权力在国家和社会之间的一种平衡和分布，当社会的形态和结构开始发展变化，原有的平衡就被打破，国家和社会之间的权力分布就开始重新调整。

不难看出，卡梅伦的“大社会”，其核心理念就是分权、开放，其权力的重新分布，是从国家流向社会。

美国社会虽然没有出现“大社会”的提法，但我们也能观察到一些类似的动态和迹象。

为了一份小小的白宫访问记录，一个草根组织可以直接向总统叫板，最后总统被迫让步妥协；因为大众的反对，一个又一个的中央数据银行计划反复搁浅；一次矿难可以引发全国性的问责，甚至不需要政府的解释，复杂的事实、曲折的真相可以在短时间内变得一清二楚、纤毫毕现，在自发产生的网络问责运动当中，政府已经从过去的裁判者变成了“被裁判者”。

除了这些微观层面的细微征兆，我们还能从美国最近发生的几次社会运动当中，看到权力结构正在发生变化的明证。

几乎在奥巴马入主白宫同时，美国兴起了茶党运动。2011年9月，美国又发生了占领华尔街的运动。这两次运动的出发点各不相同，前者最初是抗议税收过多，后者的直接诱因则是失业率太高。但这两次运动，都是网络推动的公民自由表达行动，和过去的社会运动相比，其表达方式和组织机理已经明显不同。

两个运动都没有明确的领导者或发言人，其组织结构不是“垂直”层级制的，而是“水平”的，完全是大众组织、大众参与，高度“去中心化”。当世界著名的思想领袖齐泽克来到抗议现场发表演讲的时候，他甚至都没有一支话筒。

两次运动的主题也是多中心的，茶党运动反对政府资助产业、支持政府透明，“占领华尔街”则从政治到经济，无所不包，他们打出的标语和口号有：

“建立一个好的社会，我们并不需要华尔街，也不需要政治家。”

“资本主义忘记了生活是社会性的。”

“99％的人不能再容忍1％的人的贪婪与腐败。”

这些口号，表面上是对经济不公表达愤怒、对政治系统感到挫败，其背后传达的则是权力结构要“去中心化、水平化、社会化”的诉求。齐泽克在演讲中更是清晰地发问：“怎样的社会组织才能够取代资本主义？我们希望拥有什么样的新领袖？”

还有评论一针见血地指出，这两场运动，都是美国制度失败的表

现。

美国社会现有的治理制度，从组织上来说，还是层层授权的纵向分级制，从管理方法上看，依然是以环环相扣的流程管理为基础。联邦政府是全美最大的信息生产、收集、使用和发布的单位，可谓信息社会的信息中枢，但近百年来，其层级制的治理结构并没有发生本质的变化。

但在大数据的时代，基于层级的社会控制手段将不再适用，基于流程的管理方法也有很大局限性。

这是因为，在大数据时代，信息和数据是最重要的资源，一旦信息自由、数据开放，就意味着信息和每一个公民之间都是等距的，而且中间没有层级的过滤。数据的开放和流动，就代表着知识的开放和流动、代表着权力的开放和流动，这种开放和流动，是多中心的、水平的。在这个时代，社会的主体结构将从“分层”转向“结网”，在网状传播力量的不断冲击下，个人的主体价值将得到前所未有地张扬，集中在政府的权力将开始分散，权力的最终流向，是社会、是大众、是一个个独立的公民个体。

大数据时代正在呼唤下一波社会化的浪潮：一个更开放的社会，一个权力更分散的社会，一个网状的大社会。

美国社会正在发生这种变化。

这种社会变化，当然是一个巨大的挑战。奥巴马开放数据、开放政府的举措正是试图迎接这个挑战，但这仅仅是一个开始。要应对好这个挑战，美国政府无疑还需要更多的制度创新。

注释

01 Big data: The next frontier for innovation, competition and productivity, McKinsey Global Institute, May 2011

02 英语原文为："Political technology determines political success."—Morton C. Blackwell

03 *Competing on Analytics: the New Science of Winning*, Thomas H. Davenport, 2007。此书已有中译本《数据分析竞争法：企业赢之道》（商务印书馆出版）。

04 本章所提到麦肯锡公司的研究，均指其于2011年5月发布的研究报告：《大数据：下一个创新、竞争和生产率的前沿》（Big data: The next frontier for innovation, competition and productivity, McKinsey Global Institute, May 2011）。

05 Big data: The next frontier for innovation, competition and productivity, McKinsey Global Institute, May 2011

06 孙子这句话的意思是，首先要保证自己处于最优的状态、做到不会被对方所战胜，然后等待对方犯错、伺机战胜对方。

07 Competing On Analytics, Harvard Business Review, Thomas H. Davenport, January 2006

08 《数据分析竞争法：企业赢之道》（商务印书馆出版，2009年8月），托马斯·达文波特，第8页。

09 A Data Explosion Remakes Retailing, Steve Lohr, *New York Times*, Jan 2, 2010

10 英语原文为："It is unfortunately none too well understood that, just as the State has no money of its own, so it has no power of its own. All the power it has is what society gives it, plus what it confiscates from time to time on one pretext or another; there is no other source from which State power can be drawn."—*Our Enemy, The State*, Albert Jay Nock

11 英语原文为："The transition from an industrial age to a post-industrial or information age has been discussed so much and for so long that we may not have noticed that we are passing into a post-information age."—*Being Digital*, Nicholas Negroponte

12 英语原文为："This period of history has been called the Information Age because it makes available instant access to knowledge that would have been difficult or impossible to find previously. I would argue that we are way beyond this; we're at the beginning of a new era: the New Information Age."—New Information Age, Techcrunch, Vivek Wadhwa, April 10th, 2011

13 英语原文为："You can call it liberalism. You can call it empowerment. You can call it freedom. You can call it responsibility. I call it the Big Society."—David Cameron, July 2010

# 第十三章　大变革*

## 微学位：教育领域正在发生的革命

——《南方都市报》“信息社会”之涂子沛专栏　2012年5月25日

这个月初，哈佛大学与麻省理工大学共同宣布，将投入6000万美元，开发一个在线教育平台。届时，不仅两个学校的授课过程免费向全世界开放，这个平台本身，也以开源的形式，向其他大学和教育机构免费开放。

而几个月以前，斯坦福大学已经把“在线教育”变成了硅谷最热门的话题。其计算机系的Ng教授，把自己的一门课放到了互联网上，结果全球有十几万人注册，这些人，除了在网上听他的实时讲授，还和斯坦福大学的在校生做同样的作业、接受同样的评分和考试。最后，有几千人完成了这门课程。Ng教授看到了其中的商机，在硅谷拉到了1000多万的投资，成立了在线教育公司Coursera。更早一些，谷歌无人驾驶汽车的发明者、大名鼎鼎的斯坦福教授Thrun，也离职创办了一家在线教育网站Udacity。随着类似的公司如雨后春笋般冒出，普林斯顿大学、加州大学伯克利分校、宾夕法尼亚大学等等知名学府都在近期宣布加盟在线教育，逐步向全世界开放自己的课程。

当然，这也不是大学的专利。我所在的宾夕法尼亚州，从小学到高中，都建立了在线学校，选择参加在线学习的学生，可以免费获得州政府配发的一台笔记本电脑、一台打印机和扫描仪，作为学习的工具。这是因为他放弃了公立学校的学位，节省了公共教育的开支。目前，宾州

的在线学校Cyber School，已经招收了几千名中小学生。

之所以获得教育界和投资界的一致青睐，是因为现在的在线教育，已经不仅仅是一个镜头、一段远程录相这么简单，而增加了“互动”、“互助”、“行为评价和诱导”等等新的技术和色彩。例如，师生可以随时对话，计算机能够记录你在一张幻灯片上停留的时间，判别你在答错一道题之后有没有回头复习，统计你在网上提问的次数、参与讨论的多少，然后在这个基础上，对你的学习行为进行诱导和评价。此外，学习者还可以结成互助小组，甚至互相批改、评价对方的作业和报告。哈佛大学和麻省理工学院的新网站还宣布，对于完成一门课程的学生，只要缴纳极少的费用，就可以获得一个认证，这个认证并不是真正的学位。在这个新生事物层出不穷的时代，我们不妨称之为“微学位”。

“微学位”等新概念、新技术的出现，我认为，将给教育领域带来一场革命。它的影响，不仅立竿见影，而且相当深远，会波及整个社会。

1946年，人类发明第一台电子计算机。如果以此为起点，那人类社会进入信息时代已经有半个多世纪。信息时代的特点是信息流动的速度加快、传播的广度加大，信息垂手可得、无处不在。但信息不是知识，信息必须经过加工、组织、整理才能成为知识，即将到来的在线教育浪潮，把系统整理的知识以课程为载体推上了互联网，这将帮助我们在信息时代的基础上再迈进一步，进入真正的知识时代。知识时代的最大特点，是知识无所不在、随处可得。

学校，曾经是最重要的教育资源，好的学校，更是稀缺性的资源。由于在线教育的普及，在不远的将来，名校将人人可上，教育资源匮乏的问题将得到缓解。但在线教育对个人的重大意义，还不仅仅是教育机会的增加，更是学习方式的改变。在传统的学校，我们是接受教育，但

在网络学校，人人都可以各取所需，随时随地、按照自己的节奏和计划来开展学习。例如，高中生可以尝试体验大学的课程，离开了校园的人，也可以登陆在线空间，再和在校生一起听课。通过在线教育，学习可以成为一个高度个性化的主动建构过程，终身教育也将成为普遍的现实。更多更好的教育资源、主动学习、终身教育，这些都是教育工作者探讨多年、孜孜以求的梦想。

在线教育引发的不仅仅是个人学习的革命，更会改变整个教育行业的生态。其影响大小，一时难下定论，我们不妨通过多问几个问题来催生自己的想象。以大学为例，通过把课程变成一种网络服务，而且公开竞争，越来越产品化，大学作为一个知识传播中心的作用是不是可以发挥得更好？此外，在线教育可以帮助大学突破围墙和座位的限制，名牌大学会不会因此更受追捧、其品牌效应会不会更加突出？还有，一流的大学不惧把自己的课程推上网和全世界的学习者共享，是不是因为它们还有更核心的竞争力？这个竞争力又是什么呢？是不是“创新”？顺着这个思路去想，不难得出结论：真实世界的大学会把更多的时间和精力倾斜到新知识的发现和生产上，而在线教育将承担更多的知识传播功能。

再有，学生和学校的关系是不是也会发生变化？过去，是学生争学校；将来，学校抢学生的现象可能会更加突出。一流的大学将成为一个创新型人才的集聚地，普通的大学可能要通过更好的行政管理和服务来吸引学生。再设想，如果你是用人单位，面对一名持有名牌大学“微学位”的候选人和一名三流大学的毕业生，你最后会雇用谁？“微学位”的产生，会不会引起教育行业一种新的竞争、重新布局、甚至洗牌？

哈佛大学在线教育的负责人认为，目前这场在线教育的浪潮，是印刷术发明之后，教育领域面临的最重大的变革。对中国的教育领域而言，如何合理规划、用好资源，是挑战，但更是福音，这值得我们全社

会、特别是教育部的官员、教育一线的工作者以及企业家的思考、研究和关注。

# 大数据：再谈教育领域正在发生的革命

——《南方都市报》“信息社会”之涂子沛专栏　2012年6月12日

5月25日，我发表的专栏文章《教育领域正在发生的革命》引起了一些讨论。有读者问，哈佛大学和麻省理工学院正在建设的在线教育平台及其课程，为什么要向全世界的大学和学习者免费开放？我们不妨做个补充探讨。

软件平台向全世界免费开放，这是开源软件的做法，由来已久，其背后的理念是通过开放式的协同合作，才能打造最好的平台、为社会进步贡献最大的价值。但课程向全世界的学习者免费开放，确实是这两个大学的独特之处。因为收费的做法，正是同类网站赢利的主要手段。

按理说，以哈佛和麻省理工的名气，即使收点钱，全世界的学习者也会趋之若鹜。他们之所以决定免费，有其特殊的目的。我在上篇文章中提到，新一代的在线学习平台，和过去的重要区别，在于其增加了行为评价和学习诱导的成分。向全世界开放，让最多的学习者在上面学习、使用，哈佛大学和麻省理工就可以收集最多的数据，从而研究世界各国学习者的行为模式、打造更好的在线平台。例如，通过记录鼠标的点击，可以研究学习者的活动轨迹，发现不同的人对不同的知识点有何不同的反应，用了多少时间，哪些知识点需要重复或强调，哪种陈述方式或学习工具最为有效。记录单个个体行为的数据似乎是杂乱无章的，但当数据累积到一定程度时，群体的行为就会在数据上呈现一种秩序和规律。通过分析掌握这种秩序和规律，未来的在线学习平台才能弥补没有老师面对面交流指导的不足。

这个平台到底能够收集多少数据呢？以麻省理工学院开放的一门代号为6.002X的电子电路课程为例，其校长总结说，仅仅注册这一门课的人，就比整个麻省理工所有在世校友的总数还要多。窥一斑而知全豹，这个平台最终会有多少数据，也就可想而知了。

不难看出，未来的在线教育平台之所以强大，是因于其能收集、分析、使用大量的数据。数据，是对信息的记录。数据的激增，意味着人类的记录范围、测量范围和分析范围在不断地扩大，也意味着知识的边界在不断地延伸。教育领域正在发生的这场革命，其深厚的技术背景，就是由于信息技术的进步，人类收集、存储、分析、使用数据的能力实现了巨大的跨越。这种现象，也被称为“大数据”（Big Data）。

大数据将对人类社会发生的影响，可以说难以限量。以行为评价和学习诱导为特点的在线教育平台，只是这个大潮在教育领域掀起的一朵浪花。人类社会的知识，可基本分为两个大类：自然科学和社会科学。自然科学的研究对象是物理世界，讲的是“精确”，卫星上天、潜艇下海，差之毫厘，就会谬以千里。社会科学研究的是社会现象，如经济学、政治学、社会学，它也追求精确，但因为研究对象主要是人，导致了“测不准”，所以社会科学常常被称为“准科学”。由于近年来信息技术的进步，数据的累积，个人在真实世界的活动得到了前所未有的记录，这种记录的粒度很高，频度在不断增加，为社会科学的定量分析提供了极为丰富的资源，从而可以测得更准、计算得更加精确，有科学家相信，借助大数据的推力，社会科学将脱下“准科学”的外衣，全面迈进科学的殿堂。

对此，麻省理工学院的Brynjolfsson教授曾经形容说，大数据的影响，好比几个世纪之前人类发明的显微镜。显微镜把人类对自然界的观察和测量水平推进到了“细胞”的级别，带来了历史性的进步和革命。而

大数据，将成为我们观察人类自身行为的“显微镜”，这个新的显微镜，将再一次扩大人类科学的范围，推动人类知识的增长，引领新的经济繁荣。

当然，除了学习者的行为数据，在教育领域这场革命当中扮演重要角色的还有课程视频的数据、教案的数据等等。随着在线平台的推广和普及，越来越多的老师将自己的授课视频放上网。美国有个叫Udemy的网站，老师可以根据自己发放视频的点击率获得报酬。今年5月份，该网站宣布，其年收入最高的老师，已经超过了20万美元，收入排行榜的前十名，每人的年度收入都超过了5万美元。不久前，美国又出现了一个新的网站，该网站提供工具让学习者对现有的视频按照自己的意愿进行编辑剪接，并重新发布，可以想象，这又会产生海量的数据，最重要的是，互联网上的眼睛，将使所有的错误都无所遁形，在无数人的编辑之下，知识传播的形式将不断优化、效率会不断提高。在数据的不断累积当中，最好的视频将会沉淀下来，成为经典的学习材料。除了课程视频，教案也成了美国在线平台的交易产品，佐治亚州的学前班的老师Deanna Jump，是销售榜上的头名状元，她的教案在TeachersPayTeachers这个网站上的成交量已经超过70万美元，她目前的月收入已达6万美元，比绝大多数老师一年的工资还高。教案的交易，甚至出现了由学校牵头组织的团购，统一发放给全校的老师参考使用。

可以想象，这些新的现象，会给教师这个职业带来巨大的冲击和改变。我们知道，无论是中小学的教师，还是大学的授课型教师（Teaching Professor，指非研究性的大学教职人员），都不是高收入的群体，行业的竞争性也普遍不强，由于局限于课堂、受众较小、薪酬不高，也吸引不到最好的人才。中国的情况如此，美国也一样。但美国这股正在兴起的在线教育浪潮，已经开始造就了一批万众瞩目的教育明星

和通过教育致富的高收入群体。这将提高教师这个职业的吸引力，增强教师这个行业的竞争性，新的人才可能涌入，旧的格局必将打破。回顾前文，一句话，无论是对学习者、还是教师，抑或学校，乃至整个教育产业，在线教育都将催生革命性的变化。

# 微学校：三谈教育领域正在发生的革命

——《南方都市报》“信息社会”之涂子沛专栏　2012年6月20日

前两篇文章，我从“微学位”、“大数据”这两个新的概念出发，阐述了教育领域正在发生的革命以及其技术背景。文章发表之后，引起了一些关注。为了更好地理解和应对这场教育领域的革命，我现在把它放到历史的纵轴上，再做一番考察和讨论。

教育活动的本质，是知识的传递。受限于技术水平，人类最早的教育形式，无论东方还是西方，都是以个人关系为基础的，即通过师生之间口耳相传的“传、帮、带”，完成知识的传递。这可以称为“学徒制”或“私塾制”，不难想象，这个时期的师生比例较小，教育主要依托个人和家庭来完成。

印刷术的发明，为知识的传递提供了一种新的载体：书籍。随着书籍的普及以及人类社会向工业社会的迈进，发生了第一次教育革命：现代意义上的学校出现了。而且，建立学校，慢慢演变为一种有组织的国家行为：青少年按年龄大小排序，以老师为中心、以课程讲授为形式、共同接受教育。学校的建立，使“学徒制”时代的“师徒弟子”关系转变为“师生”关系，解除了两者之间可能存在的人身依附和压迫，也打破了学徒制的师生小比例。教育，开始成为一种大规模的批量生产过程。

虽然提高了知识传播的效率，但学校也存在着一些明显的不足。例如，一对多的讲授，批量生产，无法做到对每一个学生都提供个性化的、持续性的指导和评价。又如，虽然学校成立的目的是培养学生，但老师却持有权威的地位，学生的学习，本质上还是一个按计划实施的被

动过程。

计算机发明之后，人类社会开始进入信息时代。随着信息技术的不断进步，知识的传播有了新的载体：视频。我在前篇文章中谈到，近年来，各种以课程为载体的视频，在不断地产生、累积、淘洗和沉淀，可以肯定，通过互联网上的编辑和合作，不远的将来，全世界的每一个领域、每一门课程都会出现一些经典的视频，这些视频，将取代老师在课堂上的讲授，成为最受欢迎的学习材料。换句话说，未来大部分的老师，可能会是课堂的组织者和学习的引导者，而不是知识的讲授者，其权威性将下降，师生关系将面临改变。

更重要的是，这些最好的教育资源，绝大部分都将免费。我在上篇文章中提到，哈佛和麻省理工之所以免费，理由之一，是因为“大数据”的技术背景，他们需要收集最多的行为数据，以研究开发最好的在线教育平台，并强化其学校的品牌效应。但这只是硬币的一面，对学习者免费并不代表他们无利可获。例如，他们可以通过“微学位”的认证来收费。让我们做一个保守的估计，假设一门课程的“微学位”认证费用为100美元，每一门课程每一年在全世界的范围内接受10万人的认证，以全校100门课程计，年收入将会高达10亿美元。你也许还会问，那仅仅是哈佛大学，其他一些在线教育网站，也能够免费不成？答案恰恰是“完全可能”，例如我第一篇文章中曾经提到的由斯坦福名教授Thrun创办的在线教育网站Udacity，就宣布免费，但这也不代表他不挣钱。其商业模式，是将1％学习成绩最好的学生直接输送给全世界最好的公司，从中收取中介费。换句话说，该网站不仅为全世界的学习者提供免费的学习平台，还包找最好的工作。这会在学习者当中激起怎样的竞争？对传统大学的威胁又何在？稍做冷静分析，相信你会对其中暗含的商机和变化倒抽一口冷气。

也就是说，对个人来说，知识的获得和传播，未来的边际成本将趋近于零。但这却还不足以构成这场革命。让我们回头看看：早在1980年代初，计算机就走进了学校和课堂，远程教育的概念就被提出，视频录像就开始传播，那个时候，就有人兴奋地预言：信息技术要给教育领域带来一场革命！但30年过去，信息技术确实在工业、农业、商业等领域催生了诸多革命，教育领域的革命为什么姗姗来迟、直到今年才发生？

其中的关键，是近年来，随着信息技术的进步，其“交互性”终于结出了丰硕的“正”果：以社交媒体为特征的Web2.0技术已经把人与人之间通过网络进行的交流互动推向了登峰造极的应用。教育理论认为，学习的本质是大脑对信息的一种加工，而有效的加工来自于有效的情境互动。回到我在前两篇文章中讨论的“学习行为评价和诱导平台”，现有的信息技术可以通过网络、视频和智能软件，为每一个学习者搭建起一个学习情境，像人一样对学习者的学习行为提供持续性的诱导、评价和支持，例如发出提示，建议、指出学习者的错误，帮助其形成科学的学习方法和习惯等等。除了智能软件的诱导和评价，学生还可以和其他学习者交流协同，有效地寻求帮助。美国已经出现了提供专门辅导的家教网站，学习者可以在网站上列明的专家中按课目寻找自己所需要的辅导者，在10分钟的视频交流之后，辅导服务开始按分钟计费。也就是说，学习者和辅导者可以通过很低的成本、在很短的时间内实现有效对接，构造一对一的学习情境。以上信息技术能够营造提供的种种情境，我们不妨称之为“微学校”。加上普适计算的浪潮，计算机和网络将无处不在，“微学校”将可以随时随地搭建。再想一想，因为“微学校”的存在，未来的学生可能是在家里学习，而到学校去，则主要是做作业、答疑，现有学校的运作模式可能面临颠覆。

总之，和印刷术相比，现代信息技术不仅仅创造了新的知识传播载

体，还可以为学习者在学校之外搭建一个有效的学习情境，使个人可以跳出学校的约束，自由地选择自己需要的学习方式和内容。可以预见，由于“微学校”的兴起，现有学校的教育功能将在这一场革命当中大幅衰减。教育的责任将从学校时代的国家，再度回归到学徒时代的个人，不同的是，这是螺旋性的上升，个人将对自己的学习和教育享有更多的自由、承担更大的责任，学习将彻底变成一项自我可以主导并完成的任务。这，也代表着人类本身在教育领域的一种巨大解放。

这种解放，是全社会的福音，对社会发展的正面作用不可限量，但同时也是挑战。在学校的衰减中，我预计，受到最大冲击的将是高等院校，因为它们涉及昂贵的学费以及曾经含金量最高的大学学位。在一段时间的共存之后，新兴的在线教育将最终冲垮它们中的绝大部分。未来幸存的大部分高等院校，将成为新知识的生产中心，而不是旧知识的传播中心。还有一小部分，可能要借助良好的管理和其他增值服务来吸引学生。

对中国的教育领域而言，这场革命或许还有额外的意义，那就是对现有体制的行政化和官僚化也有一定的瓦解作用。我的思考，难免挂一漏万。但仅就以上讨论而言，时下正在发生的教育革命已经值得引起全社会的重视和讨论。

* 自本书第1版出版后，大数据在社会诸多领域催生了更多变革，本章从教育领域探讨大数据的应用，并以此管窥“大数据”引发的重大变革。本章收自作者在《南方都市报》的专栏文章。

# 尾声　挑战中国：摘下“差不多先生”的文化标签

而最大的毛病，则是西欧和日本都已以商业组织的精神一切按实情主持国政的时候，中国仍然是亿万军民不能在数目字上管理。……一旦某一个国家能在数目字上管理，到底使用何种数字，尚可以随时商酌，大体上以技术上的要求作主，不必笼统的以意识形态为依归了。[01]

——黄仁宇（1918-2000），美籍华裔历史学家，1991年

在论述完大数据时代的趋势以及这个时代给个人、企业和社会带来的诸多挑战之后，本书应该可以画上句号了。但在结束之前，作为一名中国人，还是按捺不住，要絮叨几句我们中国在这个大时代当中所处的位置。

数据表明，今天的中国，是一个人口大国、互联网大国、手机大国，但却恰恰还不是一个数据大国。2011年，麦肯锡公司以2010年度各国新增的存储器为基准，对全世界大数据的分布作了一个研究和统计，中国2010年新增的数据量约为250拍，不及日本的400拍、欧洲的2000拍，和美国的3500拍相比，更是连十分之一都不到。

全世界大数据的分布




说明：新数据的存储量按年度新增的存储器容量估算。[02]（数据来源：IDC Storage Report；Mckinsey Global Insititute Analysis）

和美国相比，中国拥有数据量的绝对值较小，这在情理当中。本书第三章曾谈到，美国的联邦政府，是一个数据帝国，它的数据主要有三个来源，一是业务管理的数据，二是民意社情的数据，三是物理环境的数据。这三种数据的积累，并不是一蹴而就的，在其漫长的发展过程中，有其重要的里程碑，例如1940年罗斯福引进的民意调查、1962年启动的海浪监测计划和1973年诞生的最小数据集。

而中国类似进步的取得，都是进入21世纪之后才发生的事情。2003年，中国开始着手制定医疗系统的最小数据集，3年之后，中国卫生部出台了第一版中国医院最小数据集的标准。也是在2003年，中国创立了第一个全国性的大型社会调查项目，开始对社会的发展和变迁进行全方位、综合性、纵贯性的问卷访谈调查。这个调查叫做“中国综合社会调查”，是由中国人民大学发起的，中国人民大学随后还按照国际标准成立了“中国社会调查开放数据库”，向全社会开放调查的结果和数据。

2006年9月，在几经周折之后，国家统计局正式成立了社情民意调查中心，这是中央政府第一个、也是目前唯一一个专司社情民意调查的单位；至于对物理环境数据的采集，更是2010年前后才见到若干零星的报道。[03]

因此，从历史发展的角度来看，中国拥有的数据量较小，可以理解；但另一方面，立足现实，中国的人口、互联网的用户、手机的持有量都位居全世界第一，2010年全年新增的数据量却尚不及美国的十分之一，也大大出人意料。

2011年，中国拥有4.8亿的互联网用户，几乎是美国的两倍；拥有近9亿部手机，是美国的3倍。互联网和手机，都是产生数据的重要来源。在中国，存储器的价格也已经和国际接轨、相对便宜。在认真考察了这些因素之后，就不难发现，中国缺乏的不是可供收集的数据，也不是收集数据的手段，而是收集数据的意识。

中美两国政府关于收集三种主要数据的对比

| 数据种类 | 收集对象 | 收集方法 | 美国的里程碑 | 中国的进步 |
| --- | --- | --- | --- | --- |
| 业务数据 | 下级部门和各类社会组织 | 以基层上报、被动接收为主 | 1973 年卫生部引进最小数据集 | 2006 年卫生部制定了第一套最小标准数据集 |
| 民意数据 | 单个公民或组织 | 投入人力财力、主动收集 | 1940 罗斯福总统引进民意调查 | 2006 年国家统计局正式成立了社情民意调查中心 |
| 环境数据 | 自然环境、动植物、物体 | 以传感器自动采集为主 | 1962 年的海浪监测计划 | 2010 年起物联网的技术开始在铁道部、交通部得到应用 |

全世界互联网用户最多的**10**个国家

|  | 国家 | 2011年人口（估算） | 2000年互联网用户数 | 2011年互联网用户数 | 网民数占国家人口比例 |
|---|---|---|---|---|---|
| 1 | 中国 | 1,336,718,015 | 22,500,000 | 485,000,000 | 36.3% |
| 2 | 美国 | 313,232,044 | 95,354,000 | 245,000,000 | 78.2% |
| 3 | 印度 | 1,189,172,906 | 5,000,000 | 100,000,000 | 8.4% |
| 4 | 日本 | 126,475,664 | 47,080,000 | 99,182,000 | 78.4% |
| 5 | 巴西 | 203,429,773 | 5,000,000 | 75,982,000 | 37.4% |
| 6 | 德国 | 81,471,834 | 24,000,000 | 65,125,000 | 79.9% |
| 7 | 俄罗斯 | 138,739,892 | 3,100,000 | 59,700,000 | 43% |
| 8 | 英国 | 62,698,362 | 15,400,000 | 51,442,100 | 82% |
| 9 | 法国 | 65,102,719 | 8,500,000 | 45,262,000 | 69.5% |
| 10 | 尼日利亚 | 155,215,573 | 200,000 | 43,982,200 | 28.3% |

说明：从2000年到2011年，中国互联网的用户数增长了20多倍，已经雄居世界第一。（数据来源：Internet World Statistics，Miniwatts Marketing Group）

全世界手机使用量最多的**10**个国家

|  | 国家 | 手机数量 | 统计截至日期 |
|---|---|---|---|
| 1 | 中国 | 951,600,000 | 2011年10月 |
| 2 | 印度 | 884,371,296 | 2011年11月 |
| 3 | 美国 | 310,866,000 | 2011年6月 |
| 4 | 印度尼西亚 | 250,100,000 | 2011年5月 |
| 5 | 巴西 | 242,200,000 | 2011年12月 |
| 6 | 俄罗斯 | 224,260,000 | 2011年7月 |
| 7 | 日本 | 121,246,700 | 2011年6月 |
| 8 | 巴基斯坦 | 108,894,518 | 2011年6月 |
| 9 | 德国 | 107,000,000 | 2009年 |
| 10 | 尼日利亚 | 90,583,306 | 2011年2月 |

（数据来源：维基百科List of countries by number of mobile phones in use条目）

手机数据的应用：下一波数据创新的中心

手机是大数据时代海量数据的一个重要来源。普通手机定位（Mobile phone tracking）、基于智能手机的地理位置服务（Location Based Services）将成为下一波数据创新的核心应用之一。原因在于，手机和用户如影随形，可以记录一个人的地理位置，这意味着无数的商机。例如：

·通过手机为驾驶人员提供智能导航、避开交通拥堵·通过手机支付高速公路的过路费

·通过手机追踪人员和车辆的旅行安全、提供紧急救助

·通过手机在社交网络中实时播报用户的地理位置，与朋友和商家共享

·保险公司可通过分析手机用户地理位置的变化和行为特点，制订车辆保险费的收取方案。著名的咨询公司埃森哲正在开发策划一个新的项目以量化高速公路上广告牌的效果，其核心手段就是分析手机数据。通过手机产生的地理位置数据，可以计算每天有多少行人经过某个特定的广告牌、从什么方位经过、什么时候经过，行人中有多少当地的、又有多少外地的等等。通过这些数据，确定广告的效果，为不同用户找到最佳的广告地点和时间。

大数据时代的这些新应用和新的商业模式将影响改变全球每一个人的生活。

中国人数据意识的淡薄，由来已久，甚至可以称之为国民性的一部

分。

19世纪中叶，中国的大门被西方的坚船利炮打开。由于科学落后、文化封闭，此后的百多年来，中国人饱受屈辱。在落后、挨打的痛苦挣扎中，中国人开始慢慢睁眼看世界，反省自己与西方的差距，在这个过程当中，也有思想先贤对国民“数据意识”方面的问题进行反省、批判和鞭笞。

中国近现代著名的思想家胡适就对中国人“凡事差不多、凡事只讲大致如此”的习惯和作风深感忧虑。1919年，他写下著名的《差不多先生传》，活灵活现地白描了中国人取道中庸、不肯认真、甘于糊涂、拒绝精准的庸碌形象：

> 你知道中国最有名的人是谁？
>
> 提起此人，人人皆晓，处处闻名。他姓差，名不多，是各省各县各村人氏。你一定见过他，一定听过别人谈起他。差不多先生的名字天天挂在大家的口头，因为他是中国全国人的代表。
>
> 差不多先生的相貌和你和我都差不多。他有一双眼睛，但看的不很清楚；有两只耳朵，但听的不很分明；有鼻子和嘴，但他对于气味和口味都不很讲究。他的脑子也不小，但他的记性却不很精明，他的思想也不很细密。
>
> 他常常说：“凡事只要差不多，就好了。何必太精明呢？”
>
> ……
>
> 后来他在一个钱铺里做伙计；他也会写，也会算，只是总不会

精细。十字常常写成千字，千字常常写成十字。掌柜的生气了，常常骂他。他只是笑嘻嘻地赔小心道：“千字比十字只多一小撇，不是差不多吗？”

有一天，他为了一件要紧的事，要搭火车到上海去。他从从容容地走到火车站，迟了两分钟，火车已开走了。他白瞪着眼，望着远远的火车上的煤烟，摇摇头道：“只好明天再走了，今天走同明天走，也还差不多。可是火车公司未免太认真了。八点三十分开，同八点三十二分开，不是差不多吗？”他一面说，一面慢慢地走回家，心里总不明白为什么火车不肯等他两分钟。

……

他死后，大家都很称赞差不多先生样样事情看得破，想得通；大家都说他一生不肯认真，不肯算账，不肯计较，真是一位有德行的人。于是大家给他取个死后的法号，叫他做圆通大师。

他的名誉越传越远，越久越大。无数无数的人都学他的榜样。于是人人都成了一个差不多先生。——然而中国从此就成为一个懒人国了。

黄仁宇，是著名的华人历史学家，他参加过抗日战争，后来在美国学习、研究、深造，成为学贯中西的一代大家。他在晚年时发表了一系列的著作探讨中西方社会的异同。他认为：

“资本主义社会，是一种现代化的社会，它能够将整个的社会以数目字管理（Mathematically Management）。”

数目字管理，即以事实为基础、以数据为核心的精确管理。

黄仁宇总结说："中国过去百多年来的动乱，并不是所谓道德不良，人心不古，也不是全部军人专横，政客捣乱，人民流离"，而是因为中国未能像西方那样实行"数目字管理"的现代治国手段。他还认为："一旦某个国家能在数目字上管理，到底使用何种数字，尚可以随时商酌，大体上以技术上的要求作主，不必笼统的以意识形态为依归了。"

一句话：中国的落后，根源之一正是缺乏以数据为基础的精确管理；而未来中国的进步，也有赖于建立这种精确的管理体系。

除了中国人自己思想家的反省和批评，西方世界对我们中国人漠视数据的特点也洞若观火。

阿瑟·史密斯（Arthur Smith）是一位美国的传教士，他于1872年来到中国，在中国的普通社区生活了54年，随后著书立说，成为名动一时的中国文化研究专家。1894年，史密斯出版了《中国人的性格》一书[04]，该书在中国社会、国际社区都产生过很大的影响，一度被公认为研究中国人最权威、最详尽的著作之一，被翻译成很多个国家的文字。

史密斯认为，中国人，是一个"漠视精确、思维含混"的民族，他在书中写道：

"中国人完全能够像其他民族一样学会对一切事物都非常精确——甚至更加精确，因为他们有无限的耐心——但我们必须指出的是，他们目前还不重视精确，他们还不知道精确是什么。如果这一看法是正确的，那么就可以有两条推论：其一，在我们考查中国历史档案时，必须考虑到中国人漠视精确这一特性。我们采用中国人所提供的数字和数量

很容易使我们自己受骗，因为他们从来就不想精确。其二，对于中国人所提供的冠以‘统计数字’以抬高其权威性的各种材料，必须留有很大的余地。”（《中国人的性格·漠视精确》）

作为一名留学生，长期在美国学习、工作和生活，我也深感到，和美国人相比，中国人确实缺乏“用数据来说话”的素养。中国的语言表达方式中“重定性、轻定量”的特点非常明显，口语中经常使用“大概”、“差不多”、“少许”、“若干”、“一些”等等高度模糊的词语。我身边的一位攻读博士学位的朋友，曾经给一位美国教授提供过一份中国菜的烹调方法，但其中关于“盐少许”、“酒若干”、“醋一勺”的提法，令这位美国教授抓耳挠腮、不知所措。我事后自问，也确实感到这种提法，完全是跟着感觉走，让人无从下手。此外，和美国数据资源的丰富、开放相比，不少中国留学生对于国内的数据匮乏有切肤之痛。写论文、做研究，如果能选一个自己熟悉的中国话题，于国于己，都更有意义。可选的话题不少，但常常因为一数难求，多数选题都逃不出夭折的命运！有些数据国内确实没有，有待收集；还有些数据水分太大，经不起推敲和检验；再有的，被有关部门贴上了“机密”的标签，平民百姓无缘相见。

黄仁宇还对中国人缺乏“数据精神”的根源做了剖析，他认为：在中国传统的学问——理学或道学当中，一直都分不清伦理之“理”与物理之“理”的区别。这两个“理”混沌不分的结果，是中国人倾向于粗略的主观定性、排斥精确的客观定量，从而养成了重形象、重概括、轻逻辑、轻数据的文化习惯。这种文化习惯，使中国人长期沉浸在含蓄、模糊的审美意识当中，凡事只能在美术化的角度来印证，满足于基于相似的“模糊联想”，止步于用逻辑来分析、用数据来证明，最终将表象上的相似，当做本质上的相同。

但欧洲在中世纪之后，就将“伦理之理”与“物理之理”这两个“理”划

分得很清楚了。

黄仁宇的看法，并不是一家之言。现代著名思想家汉娜·阿伦特（Hannah Arendt）也曾经对比过中国人和西方人在思维方法上的不同。她的结论是：西方人是“词语思维”，中国人则是“形象思维”，而形象思维不是一种基于逻辑的推理性思维。[05]

归根到底，中国人对数据的漠视，缘于一种文化上的缺欠：随意、盲目、不求甚解、理性不足。

因为这种文化上的不足，科学最终在西方国家起源，当然不足为奇，中国人100多年来只能跟在别人身后亦步亦趋，也是应得的命运。

除了数据的收集和使用，中国在大数据时代需要面对的挑战，还有数据的开放。

如果说收集数据是一种意识，使用数据是一种文化、一种习惯，那是否开放数据则是一种态度。

正如本书前文所描述的，互联网上的数据开放，其开放的对象不仅仅是一个国家的人民，而是全世界的人民。这种开放，是无法阻挡的。中国人，只要愿意，就可以登陆其他国家的开放网站，下载他们的公共数据，分析他们的财政开支细目，考察他们的房屋交易价格，甚至参与他们对公共事务、社会事务的批评和监督。即使作为外国人，在这个全球化的时代，中国人也可以感受到数据开放给全人类带来的种种便利和实惠。而这些便利和实惠，正在成为信息时代世界各国一项基本的公民权利。

由于互联网的发明，“开放”已经成为人类社会一个不可逆转、不断

加速的社会思潮。如雨果所说：你可以阻挡一支入侵的军队，但你无法阻挡一种思想。在这个浩浩荡荡、不断前进的世界大潮当中，我们将发现，中国如果不跟上，我们的处境将会越来越微妙、越来越尴尬、越来越孤立。

收集数据、使用数据、开放数据，都是大数据时代我们中国人需要一一面对的挑战。

这三大挑战，没有一个不是任重道远。但这些挑战，也是我们在大数据时代彻底摘掉“差不多先生”文化标签的重大历史机遇。如果在这个数据意义凸显的时代，我们还抓不住这些历史机遇，继续漠视数据、拒绝精准、故步自封，等待我们的，还将是一个落后的100年。

## 注释

01 本引语共分两部分，分散于作者所著的《赫逊河畔谈中国历史》（三联书店，1992年）一书的正文和“大陆版卷后琐语”。

02 具体估算方法请见麦肯锡全球研究所的报告Big data: The next frontier for innovation, competition and productivity，第103页。

03 笔者没有收集到中国政府在2010年前使用传感器采集物理环境数据的实例，因此不作结论。

04 *Chinese Characteristics*, Arthur H. Smith, 1894

05 *The Life of the Mind*, Hannah Arendt (1906-1975), Edited by Mary McCarthy, 1978

# 大事记　20世纪大萧条后美国信息开放、技术创新之路

1933　罗斯福总统发起“炉边谈话”，广播技术开始在美国政治生活中得到大规模使用

1936　盖洛普领导的美国舆论研究所在总统大选的民意调查中击败《文学文摘》，把民意调查推进到科学化的历史新阶段

1940　罗斯福首次将民意调查的方法引进到公共政策的制定过程中

1945　美联社的执行主编库珀首次提出并定义“知情权”

1946　人类历史上第一台电子计算机在美国费城问世

1948　美国报纸编辑协会成立了“世界信息自由委员会”，致力于推动信息自由

1955　IBM的两名雇员辞职创办了全世界第一家独立的专业软件公司：Computer Usage Company；
国会议员摩斯提出《信息自由法》草案，倡导制度化的信息公开；
IBM推出了第一款商用硬盘存储器，每兆字节存储量的单价约为6000美元

1958　“软件”（Software）这个概念被正式定义并在出版物中使用

1960　美国举行了历史上第一次总统候选人电视辩论，此后电视在政治

生活中得到大规模应用

1962 “圣灰星期三”的风暴促成了军民联手的“海浪监测计划”，传感器技术开始进入民用领域

1965 联邦政府第一次提出“中央数据银行”方案，计划把政府部门所有的数据库都整合联结起来；
摩尔提出摩尔定律：计算机硬件的处理速度和存储能力，一到两年将提升一倍；
约翰逊总统签署《信息自由法》

1966 联邦政府交通安全管理局开始实施循“数”管理的方法，此后交通事故死亡率不断下降；
联邦政府劳工统计局开展了美国历史上第一个国家纵向社会调查，该调查长达24年

1969 IBM宣布停止免费配送软件的做法，软件和硬件开始分开定价

1970 IBM的研究员科德发明了关系型数据库

1973 美国在医疗卫生领域制定推出了第一个最小数据集

1974 美国通过《隐私法》，该法重点防止政府滥用其保存的公民信息和数据记录

1975 美国通过《信息自由法修正案》，将信息是否公开的最终裁判权赋予法院，此后知情权得到了切实保障；
盖茨创办微软公司

1976 乔布斯成立苹果电脑公司；

美国通过《版权法》，规定政府的工作成果不适用版权保护，其数据和信息只能全民共享、免费发布；
盖茨发表《致爱好者的公开信》，软件专有的概念开始成为软件产业的主导；
美国通过《阳光政府法》，公民获得了旁听政府会议的权利

1980　为规范联邦政府的信息收集工作、减少“信息扰民”，美国通过了《纸面工作精减法》

1983　麻省理工学院的斯托曼发起自由软件运动，设立GNU项目；
耶鲁大学的塔夫特教授出版了《定量信息的视觉展示》，成为数据可视化的开山之作

1988　施乐公司的科学家韦泽提出“普适计算”，预测人类的第三股计算浪潮将是：万物皆联网，无处不计算

1989　图灵奖的主办单位计算机协会举办了第一届数据挖掘学术年会，数据挖掘的技术开始兴起；高德纳咨询公司的德斯纳为“商务智能”给出了一个被广泛接受的定义

1991　GNU项目基本完成，成功推出了免费操作系统“Linux”

1992　恩门出版《数据仓库之构建》一书，第一次给出了数据仓库的定义和构建方法

1993　每兆字节存储量的单价下降到1美元左右；
关系型数据库的发明者科德定义了联机分析及其构建方法；
克林顿提出信息高速公路计划，开始兴建覆盖全美的光纤通信网络；

司法部部长雷诺发布《雷诺备忘案》，规定政府所有的信息适用于“默认公开推定”

1994　纽约街头警察梅普尔发明设计了“CompStat”警务信息管理系统，此后风行全国

1995　尼葛洛庞帝出版了《数字化生存》一书，率先提出了后信息时代的概念

1996　美国通过《电子信息自由法》，规定政府的电子记录也属于信息公开的范围

1997　南加州大学的切诺柏教授第一次定义了“云计算”的概念，指出计算的边界将由经济的规模效应决定

1998　麻省理工学院的万维网联盟提出下一代互联网“语义网”的定义和构建方法；
开源软件的概念被正式提出并定义，开源促进会成立

1999　美国客户关系管理软件提供商Salesforce提出了“软件即服务”的概念

2000　巴尔的摩市市长奥马雷推出基于城市管理的数据分析系统“CitiStat”，随后全美多个城市效仿；
美国通过了《数据质量法》，对政府所发布数据的质量进行了规范

2001　高德纳咨询公司指出，全世界的数据开始在数量、种类、增长速度三个维度爆炸；
“9·11”恐怖袭击事件爆发，美国开始制定一系列保守的法案和计

划，司法部发布了埃斯克劳特备忘案，否定了信息的“默认公开推定”原则；
加州州政府首次应用数据挖掘的技术来打击社会福利项目中的滥用和造假

2004　联邦政府农业部启动“全国动物身份识别系统”项目，为全国的新生牲畜装配移动传感器的耳标，监控动物疫情；
联邦政府国土安全部成立了国家可视化分析中心，推动数据可视化技术在政府部门的应用

2006　亚马逊公司推出首款硬件服务的云产品“弹性云计算”（EC2）；
布什总统签署《科伯恩—奥巴马法案》，其后联邦政府数据开放网站USAspending.gov上线发布，成为美国预算开支透明的里程碑

2007　美国通过《开放政府法》，再度扩大信息公开的范围；
美国开放公共数据的推动者在加州举行了第一次专门的集会，制定发布了开放公共数据的8大标准；
《连线》杂志主编凯利和技术活动家沃尔夫创办了网站quantifiedself.com，成为自我量化运动的枢纽

2008　奥巴马当选总统，因其强大的互联网动员能力，被称为首任“网络总统”，互联网在美国政治生活中开始得到大规模的应用；
华盛顿特区政府推出“城际数据仓库”的项目，向全社会开放城市管理的公共数据

2009　奥巴马就任总统，并在上任首日签署了《透明和开放的政府》总统备忘案，其后任命了联邦政府历史上第一位首席信息官和首席

技术官；
联邦政府数据开放门户网站Data.Gov上线发布；
奥巴马宣布“白宫访客记录”将在互联网上公开发布

2010　每兆字节存储量的单价下降到1美分左右；
联邦政府的年度信息技术预算为784亿美元，是全世界最大的信息技术消费者，其拥有2094所数据中心、约10000多个信息管理系统、2000个域名和24000个网站；
联邦政府举办了第一次开放政府数据的国际会议，英国、澳大利亚、巴西等十几个国家的100多名代表参加了会议；
为应对大数据的挑战，联邦政府正式提出了“大数据战略”

2011　美国信息产业的巨头EMC宣布建立“数据英雄奖”，联邦政府首席信息官昆德拉获得首届大奖；
联邦政府正式发布《联邦政府云战略》，在政府部门全面推行“云计算”；
美国和巴西共同发起成立“开放政府联盟”，并发布了《开放政府宣言》；
奥巴马在白宫召开了第一次推特会议；
共和党举办了美国历史上第一次总统候选人推特辩论；
美国气象局在全国2000多辆客运大巴上装备传感器，实时收集沿途地点的温度、湿度、光照度等天气数据；
联邦政府宣布将与印度政府合作，把Data.Gov打造成开源平台，向全世界开放

2012　美国政府启动“大数据研发计划”，投入2亿多美元推动大数据提取、存储、分析、共享、可视化等领域的研究，并将其与历史上对超级计算和互联网的投资相提并论

# 译名表　本书涉及的美国政府和民间组织机构

ACM：Association of Computing Machinery　计算机协会

ACM SIGKDD：Special Interest Group on Knowledge Discovery and Data Mining　计算机协会知识发现和数据挖掘小组

ACLU：American Civil Liberties Union　美国公民自由联盟

AHA：American Hospital Association　美国医院协会

AIPO：American Institute of Public Opinion　美国舆论研究所

ASNE：American Society of Newspaper Editors　美国报纸编辑协会

BJA：Bureau of Justice Assistance　国家司法援助局

BLS：Bureau of Labor Statistics　劳工统计局

BPD：Bureau of Public Debt　公共债务局

CBS：Columbia Broadcasting System　哥伦比亚广播公司

CCUSA：Chamber of Commerce of the USA　美国商会

CDC：Central Disease Control　美国疾病控制中心

CIA：Central Intelligence Agency　中央情报局

CMS：Centers for Medicare and Medicaid Services　联邦医疗保险和补助

中心

CPI：Center for Public Integrity　公共廉政中心

CPSC：Consumer Product Safety Commission　消费者产品安全委员会

CREW：Citizens for Responsibility and Ethics in Washington　华盛顿责任道德公民中心

DARPA：Defense Advanced Research Projects Agency　国防部高级项目研究所

DHS：Department of Homeland Security　国土安全部

DODC：National Oceanographic Data Center　国家海洋数据中心

EPA：Environmental Protection Agency　美国环境保护局

FBI：Federal Bureau of Investigation　联邦调查局

FCC：Federal Communications Commission　联邦政府通信委员会

FDA：Food and Drug Administration　美国食品与药品管理局

FHA：Federal Highway Administration　美国高速公路管理局

FSF：Free Software Foundation　自由软件基金会

FWS：Fish and Wildlife Service　美国野生动物和渔业局

GSA：General Service Administration　联邦政府总务局

IAO：Information Awareness Office　信息触角办公室

IAS：Institute for Advanced Study　普林斯顿大学高等研究院

MSHA：Mine Safety and Health Administration　矿山安全健康局

NCVHS：National Committee on Vital and Health Statistics　国家生命健康统计委员会

NHLBI：National Heart, Lung, and Blood Institute　国家心肺血研究所

NHTSA：National Highway Traffic Safety Administration　国家高速公路交通安全管理局

NIH：National Institutes of Health　国家卫生研究所

NIJ：National Institute of Justice　国家司法研究所

NOAA：National Oceanic and Atmospheric Administration　国家海洋与大气管理局

NPR：National Public Radio　国家公共电台

NSA：National Security Agency　国家安全局

NUL：National Urban League　全国城市联盟

NVAC：National Visualization and Analytics Center　国家可视化分析中心

NWS：National Weather Service　国家气象局

OFR：Office of the Federal Register　联邦日志登记办公室

OGP：Open Government Partnership　开放政府联盟

OIRA：Office of Information and Regulatory Affairs　信息和管制办公室

OJJDP：Office of Juvenile Justice and Delinquency Prevention　青少年犯罪预防办公室

OMB：Office of Management and Budget　行政管理预算局（早期叫预算分析局，Bureau of the Budget Analysis）

OSHA：Occupational Safety and Health Administration　职业安全与健康管理局

OSI：Open Source Initiative　开源促进会

PARC：Palo Alto Research Center　帕罗奥多研究中心

PCAST：President's Council of Advisors on Science and Technology　总统科学技术顾问委员会

PITAC：President's Information Technology Advisory Committee　总统信息技术顾问委员会

PRC：Pew Research Center　皮尤研究中心

SAMHSA：Substance Abuse and Mental Health Services Administration　精神健康管理局

UMWA：United Mine Workers of America　美国联合矿工工会

USCB：U.S. Census Bureau　国家普查局

USPS：U.S. Postal Service　国家邮政局

USSS：U.S. Secret Service　美国特勤局

W3C：World Wide Web Consortium　万维网联盟

# 关于本书数据、专有名词体例及图片版权的说明

1. 本书的数据一般取一至两位小数，并按四舍五入计。

2. 为行文简洁，本书只对重点人物或在可能引起混淆的情况下，引用人物的全名。一般情况下，外国人名的翻译只包括姓，不包括名，但英文全名一律在后继括号内注明，以便读者查对。

3. 为方便读者阅读，涉及美国相关组织机构时，大多采用英文名称的首字母简称，要查阅全称，请参考本书附录“译名表”。

4. 本书未标明出处的图表，皆为作者自己设计绘制。

5. 本书引用的大部分图片和照片，已获得作者的授权；少数未获授权的，欢迎作者见书后，与本书的作者或出版社联系。

# 后记　搭建“大数据”的世界

2008年12月，我刚从卡内基梅隆大学毕业，就碰上了美国的金融危机，被卷进了失业大军。

两年多的留学生活，既紧张又辛苦，结果却没工作，不免郁闷。但“宅”在家里，却有了和女儿一起玩耍、抒情咏志的时间。我开始写博客，记录自己在异国他乡的生活和感受。后来，一篇文章被中山大学的郭巍青教授读到，经他推荐，发表在2009年9月的《时代周报》上。

没想到，一个崭新的写作空间，就此豁然打开。

此后，我陆续接受了《新快报》、《南方都市报》、艾瑞网等报纸网站的邀约，写了近一年的专栏。期间，郭巍青教授、《新快报》的杨震编辑、《南方都市报》的邓志新编辑和《时代周报》的彭晓芸编辑都对我的写作有过帮助，令我受益匪浅。

但促成本书最重要的原因，是我们这个时代的迅猛发展，以及我在其中的职业经历。

1996年，作为计算机专业的毕业生，我的第一份工作是数据库程序员。后来担任了边防巡逻艇的指挥官，从事过政府部门的统计工作。但没想到，在美国的工作，又和数据、政府发生了关系。我所在的美国公司，是政府合同商，致力于为各级政府部门提供数据解决方案。我的新职位，和十几年前在中国的第一份工作极为相似，但所使用的技术已经无法同日而语。通过和一个又一个项目的“亲密”接触，我真真切切地“透视”到数据在美国政府和企业当中的重要作用。在深入了解其成

因、背景和趋势的过程中，我常常被数据的力量和美感所震撼。

我将这种新的特点和趋势概括为“大数据时代”的“循数管理”和“数据竞争”。

作为一名曾经在中国工作过10年的公务人员，我知道，这种管理方法和竞争态势还不为我们的政府、企业和大众所熟悉。

这触发我产生了一个系统的中文写作计划。

就在我决心暗下之时，我的职业发展突然峰回路转：2010年下半年，被公司提拔成技术部门的主管。在美国带队伍——立刻又成为一个挑战。为了做好本职工作，我无暇他顾，本书的设想，甚至专栏写作，都被迫搁置。

次年3月，在回国期间，我和东西网的张文武总编提到了这本书的构思。他听后眼睛一亮，说这是一本好书，建议我快快写出来。

他的眼神，又拨动了我心底那根弦。加上新岗位的工作慢慢驾轻就熟，写作的计划，于是重上日程。

从2011年6月起，全书历经8个月。这8个月当中，我几乎断绝了一切的社交、娱乐和休闲，投入了全部的业余时间。如精卫填海、乌鹊衔枝，每晚夜深人静之时，我就坐在书桌和电脑前，一点一点搭建“大数据”的世界。无论白天黑夜，一有空，谋篇布局、遣词造句的问题就在脑中盘旋。记得好几次，下班回家，过了家门的路口，还浑然不觉，一直把车往前开，最后绕了个大弯才调头回来。

本书的完成，首先要感谢我在华中科技大学的同窗王怡河。虽然都是理工科出身，但在校期间，我们就对文、史、哲很感兴趣，常常聚在

一起讨论国家和社会的问题。怡河博闻强记，对我的写作，给予了很大帮助。本书的每一篇章，他都是第一读者，一些精彩片段，凝聚有他的智慧。挚友张炳剑也曾阅读过部分章节，提出过很好的意见。我还要感谢KIT Solutions软件公司的总裁张小彦博士。小彦在美国工作、生活了30多年，对美国的政治、经济、文化都有深入的体察和了解。与他的日常交流和思想碰撞，是我写作灵感的重要来源。

广西师范大学出版社的曹凌志主编、王家胜编辑对本书的编辑和出版倾注了心血。戚丽艳、肖准对个别文本和图片的翻译有贡献。杨文卓、王玺等好友曾关心过本书的写作。在这里，无法列出所有我心存感激的人员。我在34岁的时候，下决心负笈海外，中山大学的周超教授、华中科技大学的梁木生教授都曾在我人生的这个关键点上对我指点、关心、鼓励良多。

除了辛苦和感激，在我的心里，还堆积了愧疚。

这8个月中，记不清有多少次，5岁的女儿奕奕拿着玩具、来到我的书房门口，轻声地问：

“爸爸，你可以跟我一起玩吗？”

我总是保持统一的口径，转过身、硬着心肠回答说：

“爸爸在工作，没时间……”

带着失望的表情，她怏怏地离开。

一个周末的下午，大雪初晴。我坐在电脑前修改最后一个章节。起身休息，踱步到窗口，忽然看见女儿和太太俩人正在给我停在院子中的车清扫积雪。奕奕的身高，刚过车头，她穿着大衣、戴着手套，站在厚

实的雪地中，一下一下挥动着小铲。隔着玻璃窗，我听不到声音，第一反应是抬脚下楼，一转念，又不忍打扰这个画面。我静静地靠在窗口，看着红色的车身慢慢从白色的积雪当中露出来，心，像被抚摸了一样，感到特别的温暖。

我知道，很多事难以周全。我期望这本书，能为他人带去知识和视野。

旅居海外的经历，让我深深感到，美国，是一个值得中国人认真学习的国家。所谓的"中国崛起"，我认为现阶段还是要证明我们有正确的学习态度和敏锐的学习能力。不要有桥不走，硬要下河摸石头。

要学习、要上路，这是中国的当务之急。我们的进步，应该以全世界的文明为基础。在这个过程中，中国需要对西方世界客观的观察者、理性的思考者和开放的分享者。

海外的华人、留学生拥有这样的历史机遇。我也愿把这作为自己的坐标，为国家的进步尽绵薄之力。

为了确保本书引用的资料和数据准确，我查对了很多资料。语言表达上，也做了反复的修改。尽管如此，肯定还有错漏和不足。感谢您的阅读，也欢迎您的意见和评点，以备再版时修订完善。

涂子沛　于美国匹兹堡

2012年2月28日

# 建言　中国如何应对大数据时代的挑战*

最近，我回国参加了一些大学的研讨会、政府部门的座谈会以及企业的培训活动，主题都是大数据，时下，国内各大报刊杂志也都在探讨这个热门话题，但我发现，中国社会对“大数据”的概念还存在一些不准确的认识甚至观念上的误区，特别是对大数据在国家层面上的战略意义估计不足，亟需深化。

# 从小数据到大数据

“大数据”是一股新的技术浪潮，也是逐步形成的历史现象，其具体是指随着信息存储量的增多，人类在实践中逐渐认识到，通过数据的开放、整合和分析，能发现新的知识、创造新的价值，从而为社会带来“大科技”、“大利润”、“大智能”和“大发展”等新的机遇。大数据概念的提出，可以追溯到1980年代，但其“数据”二字却和我们传统的理解有所不同。

传统意义上的“数据”，是指“有根据的数字”，但在进入信息时代之后，“数据”二字的内涵在扩大，它不仅指代“数字”，还统称一切保存在电脑中的信息，包括文本、声音、视频等。更重要的是，随着信息技术的进步，其数量在爆炸，特别是新媒体出现之后，数据的收集、保存、维护、使用等任务，成为横跨各个领域的现象和挑战。

大数据之“大”，并不在于其表面的“大容量”，而在于其潜在的“大价值”。有很多例子可以证明，由于新工具的出现，我们从以前的小数据当中也能发现大的价值。例如，美国把20多年的犯罪数据和交通事故数据映射到同一张地图上后惊奇地发现，无论是交通事故和犯罪活动的高发地带，还是两者的频发时段，都有高度的重合性。这引发了美国公路安全部门与司法部门的联合执勤，通过共治数据“黑点”，交通事故率和犯罪率双双降了下来。再例如，最近有学者将白宫200多年总统洗衣服的记录电子化，然后进行分析，也得出了一些新的结论。这些数据，都是地道的小数据。这说明，小数据只要在纵向上有一定的时间积累，在横向上有细致的记录粒度，再和其他数据整合，就能产生大的价值。从这个角度来看，大数据也可以理解为针对某个对象在时空两个维度上的“全息”数据。这种“全息”，在大数据的时代还表现为“多源”，即有多

个源头在从不同方向对同一个对象进行数据记录，数据之间可以互相印证。

另外，从全球数据技术投入的资金分布来看，传统的小数据仍然占据绝对的重头。据国际数据集团（IDG）统计，2012年，全球对小数据分析工具的投资为349亿美元，而对大数据分析工具Hadoop的投资仅为1.3亿美元，不及前者的1％。IDG的结论是，传统的小数据软件满足了企业和组织95％的需求。目前行业发展的最新态势，是“大”、“小”数据分析工具趋于一体化并在向“云”迁徙。

# 大数据的战略意义

大数据的意义，也远远不局限于我们当前众多新闻报道中所津津乐道的“啤酒和尿布”等通过数据挖掘、实现精准营销的故事。事实上，数据挖掘已经不是大数据领域的前沿，取而代之的是机器学习。数据挖掘是指通过特定的算法对大量的数据进行自动分析，从而揭示数据当中隐藏的历史规律和未来的发展趋势，为决策者提供参考。时下兴起的机器学习，凭借的也是计算机算法，但和数据挖掘相比，其算法不是固定的，而是带有自调适参数的，也就是说，它能够随着计算、运行次数的增多，即通过给机器“喂取”数据，让机器像人一样通过学习逐步自我提高改善，使挖掘和预测的功能更为准确。这也是该技术被命名为“机器学习”的原因。这也是大数据之所以被称为革命性现象的根本原因，因为从本质上来说，它标志着我们人类社会在从信息时代经由知识时代快速向智能时代迈进。

不妨举一两个例子，来说明大数据对社会形态的影响以及对国家战略的重要性。

今年以来，一股在线教育的浪潮正在席卷美国的教育领域，一种新型的智能学习平台正在成为高科技领域创新和投资的重点，其中不少公司已经获得了初步成功。如著名的在线教育公司Coursera，已经和普林斯顿、伯克利、杜克、香港理工等全世界30多所大学达成协议，通过其平台免费开放课程。如今这些学校的课程可以实现全球几十万人同步学习。分布在世界各地的学习者不仅可以在同一时间听取同一位老师的授课，还和在校生一样，做同样的作业、接受同样的评分和考试。一些学校看到了这种智能学习平台的价值和潜力，甚至开始投资兴建自己的独立平台，2012年5月，哈佛大学和麻省理工学院就宣布，将投入6000万

美元开发一个类似平台，并向全世界免费开放。

这种学习平台的崛起，在美国引起了广泛的关注和激烈的讨论。其中的原因，在于该平台已经不是一个镜头、一段视频那么简单，而能对学习者的学习行为自动进行提示、诱导和评价，从而弥补没有老师面对面交流指导的不足。例如，通过记录鼠标的点击，计算机能够记录你在一张幻灯片上停留的时间，判别你在答错一道题之后有没有回头复习，发现不同的人对不同知识点的不同反应，从而总结出哪些知识点需要重复或强调，哪种陈述方式或学习工具在哪种情况下最有效等规律。

不难发现，该平台之所以强大，正是因为大数据。单个个体学习行为的数据似乎是杂乱无章的，但当数据累积到一定程度时，群体的行为就会在数据上呈现一种秩序和规律。通过收集、分析大量的数据，就能总结出这种秩序和规律，然后有的放矢，对不同的学习者提供有针对性的帮助。哈佛大学和麻省理工学院之所以向全世界免费开放其学习平台，目的也是想让更多的学习者在上面学习，以收集更多的数据，有了数据，它们才能研究世界各国学习者的行为模式，进而打造更好的智能学习平台。

# 数据驱动的智能时代

前面的例子说明数据正在成为组织的财富和创新的基础，也证明大数据确实在催生一个更加智能的社会。那么，又该如何理解我们正在迈进的这个智能型社会呢？

理解这个问题的关键在于，无论是信息、知识还是智能，在我们这个时代，都是以数据为载体存在的。数据是对客观世界的记录，当我们对数据赋予背景时，它就成为信息；信息是知识的来源，当把信息提炼出规律的时候，它就上升为知识；知识是智能的基础，当电脑、网络能够利用某种知识作出自动判别，采取行动为人类服务的时候，机器智能就产生了。目前，人类记录周围世界的范围正在不断扩大，过去，我们是决定记录什么，现在及将来，我们要进入一个决定不记录什么的时代，同时数据分析的能力不断增强，这都将加速我们迈向智能时代的步伐。智能时代的特点，是无处不在的计算机和网络将像有智商的人一样为人类工作和服务。换句话说，越来越多的工作将被计算机或者机器人所代替。此外，由于精准的计算和预测，整个社会可以像无数个大大小小的齿轮轴承一样，环环相扣，齿齿吻合，日常管理通过数据更加优化，各种任务、合作可以无缝对接，社会运行的成本可大幅降低。

回到上面的例子，不难想象，这种智能学习平台将会给教育行业带来怎样的影响。学校曾经是最重要的教育资源，好的学校更是异常稀缺，由于这种智能平台的普及，在不远的将来，名校将人人可上，也就是说，如果应对得当，中国教育资源匮乏的问题将很快得到有效缓解。对个人来说，随时随地地学习、终身学习都将成为可能，例如，高中生可以尝试大学的课程，离开了校园的人，也可以登录在线平台再和在校生一起听课。这些都是教育工作者探讨多年、孜孜以求的梦想。但硬币

的另一面，是中国的教育行业要面对更加激烈的全球化竞争和挑战。过去，是学生争学校；将来，可能是学校在全球范围中争夺学生。发达国家的一流大学会挤压发展中国家普通大学的生存和发展空间，普通大学该如何来吸引生源？它们会不会因此衰落？既然最好的教学视频等学习资源都可以免费获得，教师的角色又需不需要调整？又该如何调整？这些问题，都是大数据时代催生的重大挑战。

智能学习平台只是大数据大潮在教育领域掀起的一朵浪花。毫不夸张地说，大数据将影响人类社会发展的方方面面、优化改造每一个行业，其作用难以限量。我们再以时下另外一个热门名词“智慧城市”为例。近几年来，国内外都兴起了建设智慧城市的浪潮。据国内智慧城市的领军公司神州数码董事局主席郭为介绍：目前，国内已经有60多个城市把建设智慧城市纳入了“十二五”规划，他相信，智慧城市将成为推动中国经济可持续发展的主动力。但从一个更高的角度来看，智慧城市的建设问题，其实是一个城市的大数据综合治理问题：一是要在以前没有收集数据的地方收集数据，这主要是利用物联网的技术；二是要让不同系统的数据有效对接起来，这是系统整合的任务；最后，还要利用数据可视化的技术把海量数据中隐藏的知识揭示、展现出来，让数据中的智慧能够以一种直观的形式流向城市的管理者、决策者和市民大众。也就是说，数据的收集、整合、分析、展现才是智慧城市的核心，未来的智能型城市，必将是数据驱动的城市，而大数据则相当于智慧城市的大脑。郭为还指出，智慧城市的建设，是在用信息技术解决社会治理中的难题，提高人民的幸福指数，这又证明，大数据的应用和价值，绝对不仅仅是在商业领域这么简单。

除了推进社会形态的跃进、加速企业创新，引领新的经济繁荣，我在《大数据》一书中还指出，通过开放数据，大数据还可以成为启动透

明政府的利器。这对当下的中国，现实意义毋庸讳言。也正是基于以上种种战略考量，2012年3月，美国联邦政府宣布投入巨资启动大数据的研发任务，并把大数据提到了和历史上的互联网、超级计算机一样的高度，成为国家战略。

## 政府需要做什么

一是政府机构、行业组织和大型企业要建立专门的数据治理机构来统筹数据治理的工作，例如数据治理委员会、大数据管理局等，数据治理的重点在于数据定义的一致性和数据的质量。在大数据时代，不同系统之间的数据要进行整合，因此要有统一的元数据定义，这不仅是中国而且是全世界当下都在面临的挑战。各个领域和行业的数据标准制定得好，将会起到事半功倍的效果。就单个企业而言，要认识到，未来的竞争是知识生产率而不是劳动生产率的竞争，数据分析产生的价值可能比较碎片化，分布在商业流程的各个环节，数据挖掘的投资回报也有不确定性，但企业领导必须有眼光，把数据治理的工作尽快统筹起来，为增强企业在大数据时代的竞争力做好准备。此外，数据治理机构的首长应该由组织的高层领导担任，否则标准无法推进到全局，也改善不了整个行业或组织的情况。

二是开放数据。数据增值的关键在于整合，但自由整合的前提是数据的开放。开放数据是指将原始的数据及其相关元数据以可以下载的电子格式放在互联网上，让其他方自由使用。开放数据和公开数据是两个不同的概念，公开是信息层面的，是一条一条的；开放是数据库层面的，是一片一片的。开放也不一定代表免费，企业的数据，可以以收费的形式开放。开放也是有层次的，可以对某个群体、某个组织，也可以对整个社会开放。在大数据的时代，开放数据的意义，不仅仅是满足公民的知情权，更在于让大数据时代最重要的生产资料数据自由地流动起来，以催生创新，推动知识经济和网络经济的发展，促进中国的经济增长由粗放型向精细型转型升级。

三是鼓励、扶持基于数据的创新和创业。政策扶持的传统方法，可

能是以政府为主导建立大数据产业园，对新兴企业提供办公场所等便利条件或者现金支持，这固然有效，但更有效的方式是调动全社会的力量。例如，拨款支持大数据开源社区、程序员协会等民间组织的建设，通过扶持类似的民间团体，快速推进新技术、新理念在全社会的传播和普及；再例如，以开放的数据为基础，举办应用程序开发大赛，向全社会征询数据使用、创新的意见，主办方可以是政府，也可以是企业，拿出一定的资金，奖励最优秀的应用程序，激发民间蕴藏的创新力量。

四是要在全社会弘扬数据文化。数据文化，是尊重事实、推崇理性、强调精确的文化。要承认，回望历史，中国是个数据文化匮乏的国家，就现状而言，中国数据的公信力弱、质量低，数据定义的一致性差也是不争的事实。这方面，政府应该发挥主导作用，首先在公共领域推行数据治国的理念，要认识到，在大数据时代，公共决策最重要的依据将是系统的数据，而不是个人经验和长官意志，过去深入群众、实地考察的工作方法尽管仍然有效，但对决策而言，系统采集的数据、科学分析的结果更为重要。政府应加大数据治国的舆论宣传，将数据的知识纳入公务员的常规培训体系，力争在全社会形成“用数据来说话、用数据来管理、用数据来决策、用数据来创新”的文化氛围和时代特点。

最后是要围绕个人数据安全，逐步加强隐私立法。任何技术都是双刃剑，大数据也不例外。如何在推动数据开放的同时有效地保护公民隐私，将是大数据时代的一个重大挑战。

新年刚刚拉开序幕，希望中国政府相关部门尽快制定与大数据相关的政策，出台具体的措施，从而抓住历史的机遇、推动中国社会的发展和进步。

2013年，应该成为中国的大数据年。

（2013年1月8日）

* 作者涂子沛系中国旅美科技协会副主席，在2013年两会期间，本文作为海外议案，提交至中国人民政治协商会议讨论。

# 访谈　大数据时代的“中国梦”——共识网专访《大数据》作者涂子沛

□采访人：杨传银、郑子蒙　□访谈时间：2012年7月2日

# 大数据：正在到来的数据革命

杨传银：您写《大数据》这本书的初衷在哪里？您说这不是一本纯粹谈技术的书，而是以技术为背景探讨人和社会的关系的书，这个观察和写作的角度是从什么时候开始建立的？

涂子沛：这本书应该说是时代浪潮的一个产物。我是理工科出身，大学毕业后我的第一份工作就是数据库程序员。十几年之后，我在美国的第一份工作是数据仓库程序员，从数据库变成了数据仓库，这也是我们从小数据时代进入大数据时代的明证。但更巧的是，十几年前，我在中国为政府部门开发数据管理的软件，现在在美国的工作，恰恰又是为美国的政府部门提供数据管理、分析的解决方案。这是一种“轮回”，也是一种“上升”，这种经历让我找到了一个很好的角度来观察、体验、分析两个不同国家如何解决、处理一个相同的问题。无疑，中国在数据收集、使用和管理的各个方面，和美国相比都存在差距。但我发现，最大的差距还是意识和态度。我们的传统文化，长期重定性、轻定量，重观点、轻数据。在现实生活中，数据也往往成为一个任人打扮的小姑娘，得不到应有的尊重，数据的公信力严重不足。但正在到来的大数据时代，数据的作用将前所未有地凸显，数据将成为国家竞争的前沿、企业创新的来源。这些传统和现实的不足，是我们中国社会迈进大数据时代的制约和障碍，我们应该如何应对？对这个问题的思考，催生了这本书。

这本书首先是一本信息时代的科普读本。但我想写的，不仅仅是一本科普书。1997年，我在广东省边防局开发“反偷渡信息管理系统”的软件，那时候我就明白了，要编好一个程序，就必须是商务过程的专家。

其实，技术工作永远是从业务中来、到业务中去的，技术归根到底是为人和社会服务的。这个时代，我们好比身处一列快速奔驰的列车，技术就是它的发动机，我觉得幸运的是，作为信息技术的从业者，我在这台发动机的核心部件上工作过，程序员把自己对业务的理解、自己的理念转化为软件，来改造世界。但这不够，因为发动机再强大、程序员再能干也不够，火车要开到哪里去，是政治家、企业家在掌舵领航的。

基于这些体会，我努力尝试在这本书中融合技术、政治、商业三个方面的视角。而且，这本书谈的是“数据革命”，我想任何一种革命，首先都是社会层面的事。

# 大数据时代会使不理性的人更不理性

杨传银：大数据时代有没有可能使不理性的人更不理性呢？

涂子沛：有可能。这是我下一本书计划要讨论的内容。下一本书，我想探讨大数据时代潜在的弊端。这个时代，是个数据奔流的时代，信息无处不在。一个信息开放社会给不理性的人带来的问题是，可能导致他“信息过度窄化”，例如，信息开放的时候，一个偏激的人更容易找到同类，然后放大异常，会变得更加偏激，而在一个信息流动没那么快、相对封闭的时代，偏激的人像是一个“孤岛”，他更有可能会怀疑自己。仔细观察信息时代中的人的境况，你会发现知识水平越高的人往往越顽固，这是因为他更知道怎么去获得支持自己的信息、形成自己的气场。

## 数据开放的界限

杨传银：在数据开放的过程中，政府、公司和个人可能是不一样的，这个数据开放的界限在哪里？

涂子沛：目前，数据开放已经形成了一股潮流。在各个国家数据开放的实践中，基本形成了两个界限，第一就是不危及国家安全，第二是不能侵犯公民隐私。当然，对商业组织来说，还要顾及它的商业利益。

在这几个界限之外的数据，都应该开放，而且是免费开放，其中的道理也很简单，政府收集了数据，用的是纳税人的钱、收集的也是关于纳税人的数据，自然要向纳税人开放。你不能用我的钳子、拨了我的毛，再卖给我收钱，对不对？英国社会这方面的步子迈得比美国还大，一个新的名词——公民数据权，也应运而生了。这是公共领域，而商业领域，数据开放也是一个大的趋势，我们也能看到各种各样的商业数据正在走向开放。

对个人来说，关键在于要有能力、渠道、权利去管理和自己身份相关的信息，你可以决定什么时候公开，以及用什么方式公开和自己相关的数据和信息。换言之，你个人对自己的数据必须要有控制权。因为这关系到你的隐私权。

## 中国人有不尊重数据的传统

杨传银：现在国内对数据的重视您关注得多吗？据说国内可能是对内做统计，但是对外创造数据。

涂子沛：我很关注，但总的情况是，不尽人意。我在书中详细阐述了，中国人缺乏尊重数据的传统。我们现在好一些，认识到了数据的重要性，但还是有很多问题，集中表现在，数据普遍缺乏公信力。发布数据，是一件很严肃的工作，我在书中曾经提到，2000年美国国会甚至专门通过了《数据质量法》来确保政府发布数据的质量。这个法律规定，政府发布的数据，其产生方法必须是透明的，也就是说，要说清楚，数据怎么来的，每一个数据，都要有“出生证”，说明其产生的方法，这样做了，这个数据就是可以“复制”的，这意味着，个人可以通过同样的方法，获得和政府同样的数据和结论。美国的这些经验，我认为，是值得我们借鉴的。

此外，我们还要回到刚刚谈到的开放数据，一个数据是孤立的、难分对错，一片数据才可以互相支持、互证真相，所以，要提高数据的公信力，最好的办法就是开放数据、让老百姓获得一个数据背后的一片数据。

我们社会常见的利用数据弄虚作假的问题，可以总结为几个方面：一是一叶障目，片面解释一个数据，其中也可能是别有用心、故意为之；二是人为地控制数据的收集和统计过程，使结果产生扭曲和偏差；三是暗箱操作、修改数据；最后就是无中生有、“创造”数据了，这是最糟糕的，因为假数据比没数据对社会的危害还要大。

## 中美政府信息开放的差异

杨传银：您觉得中美政府在信息开放方面有哪些差异，形成这种差异的原因在哪里？

涂子沛：最大的差别当然是我们起步较晚，美国1967年就通过了《信息自由法》，中国社会在2008年前后，才制定了《政府信息公开条例》，而且实施4年以来，总体情况还不太理想。

除了起步晚，各方面制约的力量也还不成熟。信息公开是一件很难的事情，难就难在没有人愿意主动公开！虽然一谈到透明和公开，大多数人都会认可这是正确的价值观，但一旦要自己透明、要自己公开，那公开和透明就会立刻转化为一种威胁，美国政府也不例外。它的《信息自由法》，也是经历了十多年的论证、辩论甚至激烈的争议才出台的，立法过程中，政府部门还坚决表示反对。甚至1967年美国的《信息自由法》通过之后，政府部门也是消极应对的，对于公民提出披露信息的要求，政府部门往往以各种理由搪塞、拖延。后来在公民团体、新闻记者的批评和抗议中，美国国会在1972年提出了《信息自由法修正案》。该法规定，如果政府拒绝公民的信息公开要求，公民可以提起司法诉讼，而法院才拥有信息是否公开的最终裁判权。这个法律颁布之后，才真正打开了政府信息公开的大门。

所以，信息公开，不仅要有民间的诉求，还需要新闻界以及司法力量对政府的监督，我们中国的情况，这几方面的力量都明显还火候不够。

## 中国的公民社会建设，需要大众启蒙，也需要“开路先锋”

杨传银：您提到了美国推动信息公开的过程是非常艰难的，法院、新闻界和民间在这个过程中都付出了努力，您觉得如果在中国来推动这件事可以从哪方面努力？

涂子沛：美国社会最后达成了信息公开的共识，主要原因在于它的体制，权力是需要制衡的。但我在这里更想讨论一下社会层面的努力。

公民社会的要义在于自由、平等、责任和参与。我在书中也强调，一个真正的信息社会，首先是一个公民社会。没有人的自由和平等，就不可能有真正的信息自由和平等。自由和平等是公民社会的特点，但要建设一个公民社会，这还不够，还需要我们的公民愿意承担公共责任、参与公共事务。通过意见表达，凡事发出声音，防止政府犯错误。防止我们的政府犯下错误，是我们每一个公民的责任，公民社会是一个需要公共精神、能量和想象力的社会。当然，谈到参与，还有一个机制的问题。就是老百姓能够自由地组织起来，形成团体，他们的声音才会被放大，最后形成一个防止政府犯下错误的监督力量。

在一个缺乏参与精神的社会，首要的工作是大众启蒙，也就是需要一些人来做表率、做开路先锋。我们这个社会，也总是有人愿意来做开路先锋和道德楷模的，中国社会已经出现了不少这种先锋人物。“那美好的仗我已经打过，当跑的路我已经跑尽，所信的道我已经守住。从此以后，有公义的冠冕为我存留。”我在书中引用了这段话来形容美国的国会议员摩斯，他是美国的《信息自由法》之父，他一直认为信息公开

应该立法，得罪了几任总统，也很孤立，但是他始终坚持。他的坚持和努力历经十多年，发生了很多错综复杂、戏剧性的故事，才实现了信息公开立法的目标。

我们现在特别需要这种开路先锋和道德楷模的力量。他们是少数人，但历史，常常就是由这些人改变和创造的。我愿意和大家分享我书中的这样一句话：

“永远不要怀疑：那一小部分有思想并且执著努力的公民能够改变这个世界。事实上，人类的历史从来都是这样。”（本书第239页）

## 民主的本质就是参与和妥协

杨传银：但是放在中国，您刚刚说的这些问题恰恰是中国存在的，如果要把您在书中提到的这些选择在中国落实的话，还是存在很难突破的一个问题。

涂子沛：突破的关键，还是培养公民精神的问题。大家愿意去参与公共事务，愿意去发言，情况就会彻底改变。我们可以看看古今中外的历史，专制制度为什么会存在？这种制度的存在，其实是我们大部分人理性选择的结果。专制者把你的一部分权利拿掉了，被拿掉权利的人就在想：我要是反抗，把这个权利夺回来，成本和收益各有多大呢？他一计算，发现他如果要拿回他失去的权利，和专制者对抗，他可能要失去更多，成本比收益大，理解这个问题的关键是：当一个人通过抗争从专制者那里拿回了这个权利，受益的将是一个社会的全体成员，他们都可能重新获得这个权利。也就是说，整个群体都受益，但代价却可能要少数人付。每个人一理性计算，结果不约而同都放弃了。专制者其实也在计算，他在找那个专制的“度”，把自己的专制放在一个大家可以适度容忍的框架里面，因为，他不怕一个人反抗，他怕的是大家都反抗，如果超出那个度，引起大多数人的反抗，那他就完蛋了。所以要破局，主要看有多少人参与，大家都参与，形成一个集体行动的态势，情况就会彻底改变。这也是专制者要不遗余力打压集体行动的原因，一旦有集体行动，他就把那个头头打掉，其他人一看做头头要付出这么大的代价，结果就又放弃了。正因为如此，专制制度才会存在。

杨传银：击碎这一块的可能性在哪里呢？

涂子沛：集体行动的形成，最重要的是民众的自觉，他们知道了自己必须要参与。中国人的参与感比较差，我们长期处于农业社会，农业社会是靠天吃饭的，靠血缘关系，靠地、靠自己，工业社会就不一样了，是一个大协作的社会，你必须要有公共精神、协作精神。中国人为什么公共精神、协作精神比较差，跟我们进入工业社会的时间比较短也是有关系的。

当然我们更希望体制能够建立更好的分权制衡和监督，但是如果公民社会真正成熟了，其他的事情将是一个由下而上的过程，自然而然就会发生、做到了。

我觉得广东现在就做得很好，它把非政府组织放开了，看起来这是一小步，其实是历史的一大步，因为你一放开，老百姓就可以自由结成团体，他们会自动寻求把利益诉求捆绑在一起，一旦捆绑在一起，形成一个团体了，现代信息社会有很多工具，人其实很容易连接起来的，就会形成一股力量，这股力量就会形成一种监督。

杨传银：刚才您也提到了有一些人是要当英雄来推动这个事情，另一方面来说美国政府或者资本一方是有一定的妥协的，您在书中也提到了关于妥协的问题。但是在许多环境之下这个妥协可能是很有限的，比如在极权国家里。

涂子沛：民主的本质就是妥协。一个善于妥协的社会，才是一个理性的社会。妥协的发生也是有一个临界点的，当有越来越多的人参与的时候他就必须妥协，问题是维权人士常常是孤岛，给予支援的人太少，伸出手的人太少，最后轮到自己头上就嗷嗷叫，别人再受苦受难，我视而不见，过好自己的日子就行了。所以还是缺少一个公共精神，一个参

与的精神。

## 中国的当务之急是“代表专业化”和“民主操练”

杨传银：您在书中提到了关于利益代表机制问题，在中国现在也是很显著的一个问题，其中也涉及关于结社自由的授权制度，我不知道您在国外对这块了解得怎么样？

涂子沛：我先说一下利益代表，我觉得我们可以做的一件事情是把代表专业化，这是比较务实的，一旦专业化会出现很多新的情况和气象，那个大气氛就会立刻完全不一样了。现在是兼职，一旦专业化变成全职的，提出的议案质量也会更高。

杨传银：实际上这是很系统的一个问题，提到代表专业化首先一个前提就是代表的产生问题，可能在中国人大代表的产生，我们从结果上来看代表的身份比例，官员可能是占一大部分，富人是占更大的一部分。

涂子沛：我还是觉得专业化是一个途径，一旦专业化的话，产生的途径就会不一样了，就会有更多的人想去当，会有更多的人站出来参选。

杨传银：就像您刚才说的，去年就出现了一波独立参选人大代表的热潮，但是各个地方都出现了打压的现象。

涂子沛：这种现象的出现，有它的必然性，这是一个博弈的过程，

社会的进步就是这样产生的，很多人都在做铺路石，大家来参与，这个社会才会进步。我们这个社会其实呼唤这种道德英雄，像摩斯这种，也呼唤像独立参选人这样的人，我们不仅呼唤普通的公民觉醒，我们也需要道德英雄，需要开路先锋。

杨传银：前段时间柳传志说一人一票可能把中国带进万劫不复的问题，您怎么看这个问题？

涂子沛：那个信息好像传达得不完全，有点断章取义。

杨传银：有偏差，但是也是反映了一个思潮。

涂子沛：我是这么看的，民主是需要操练的，一人一票是需要操练的。台湾其实也走过了这种混乱期，今天的台湾，我们看到他们败选之后，已经像美国的候选人一样站出来发表败选感言，说我认输，我祝贺你，这都是谦谦君子之风，是一步一步操练体会出来的，当他犯了错误之后他才知道我不该这么做，所以民主是需要操练的。操练一两届，大家都会成熟，每个人都会找到合适的方式。但是在找到之前，可能是有这样一个阶段和过程的。比如现在基层的村委会的选举，现在贿选得很厉害。我们还在这个过程当中徘徊。

杨传银：我觉得这是很荒谬的事情，一边说国民素质不高，一边又在最基层可能相对来说从文化素质上来说最不高的群体里搞直接的民主。

涂子沛：对，存在这个问题。但也要看到，在最普通的老百姓中操练，有一天它最终会成熟的，我相信老百姓是会觉醒的，当大家认真来投这一票的时候，民主制度就实现了。

# 华人社会需要政治参与

杨传银：您在美国有没有参与类似的公民运动，华人社会的政治参与度如何呢？

涂子沛：我现在担任中国旅美科技协会匹兹堡分会的主席，这是我们当地最大的一个华人组织，我们这个组织常常代表华人社区，参与美国社会的一些公益活动。

我们华人有很多优点，最大优点就是遵纪守法、勤劳肯干，但在公共生活的参与上，确实有很多不足。例如公共活动的募款，一般来说，向华人募款比较困难。这里有很多原因，不是因为华人没有钱，华人大部分都是中产阶级，而是因为我们没有这个传统。但美国人一般都乐于捐款，大部分的家庭每年都有这部分预算，你说清楚了募款的目的，捐或不捐，一般比较干脆爽快。但我们华人就不一样了，大部分华人的生存状态还是各人自扫门前雪，过好自己的小生活，其他事情不关心、也不过问。回到刚才谈到的，这可能是农业社会的传统，不仅是中国大陆，海外的华人也是这样。我们进入工业社会的时间较短，还不习惯协作介入公共生活、营造公共价值。

这是普遍现象，当然也不能绝对化，我也看到了一些好的方面。我碰到一个台湾来的人，我觉得他介入公共生活就很好，哪里有选举，他就去投票、募款，各种公益活动上也能见到他的身影，我说你为什么这么积极，他说了一句话，给我印象非常深刻，他说“我比美国人还热爱美国。因为美国人生来就是美国人，我是二次选择要做美国人的，我珍惜这种权利和参与”。认识到公共参与的重要性，其实就是认识到，现

代生活是群体性的，如果脱离群体，个人就会原子化，难以发出自己的声音，这正是专制者求之不得的。

## “美国梦”与“中国梦”

杨传银：您的书里前面提到了奥巴马和美国的一人一票，里面谈到了中国梦、美国梦的问题，您最初的理想是考公务员进入政府，也算是改变中国的一个梦想。以您的亲身体会来讲，美国梦和中国梦的差别在哪里？

涂子沛：我在书中提及了美国社会有很多外来的移民，解释了他们为什么要去美国、又如何在美国获得成功的故事，其中，有政治家、企业家，也有科学家，等等。我想，美国梦最吸引人的地方，在于这是一个开放的社会、拥有相对公平的制度，在这样一个环境中，只要通过自己的不懈努力，即使是移民，也能够获得成功。一人一票，其实代表着一种自由和平等。这是美国梦的精髓吧。

虽然我现在在海外工作和生活，内心深处，却仍然有一个中国梦，我们一定会进步，这种进步，应该以全世界的文明为基础。我们未来的发展，不仅要重科技、重创新，快步迈向信息社会，还要建设一个公民社会。因为只有在一个自由、平等的社会里，每一个人的创造力才能真正迸发出来，我们的社会才会涌现乔布斯、扎克伯格这种创新型人才。创新，才是一个国家和社会持续发展的真正动力。

这是我写《大数据》一书的初衷，也是我的“中国梦”。